DISTANCE BASED LINEAR REGRESSION MODEL AND ITS APPLICATION TO MICROBIOME ASSOCIATION STUDIES

by

Masoumeh Sheikhi Kiasri

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Applied Mathematics

Charlotte

2021

Approved by:

_____
Dr. Shaoyu Li

_____
Dr. Yanqing Sun

_____
Dr. Eliana Christou

_____
Dr. Shaozhong Deng

_____
Dr. Donna Kazemi

ABSTRACT

MASOUMEH SHEIKHI KIASRI. Distance Based Linear Regression Model and Its Application to Microbiome Association Studies. (Under the direction of DR. SHAOYU LI)

In the past few decades, pairwise distance based statistical methods have been developed to identify spatial and/or temporal clusters of disease, study the association between the dissimilarity of ecological communities and distance in geographical locations. With emergence of high-throughput technologies, pairwise distance base methods are widely used in the analysis of genetics and genomics data, especially when the data structure fails the fundamental assumptions of classical multivariate analysis, including independency and normality. However, much of existing knowledge has been around non-parametric or semi-parametric estimations usually employing permutation techniques to assess statistical significance, which are known to be computationally expensive and sensitive to the choice of permutation.

Majority of this thesis focuses on linear regression of pairwise distance matrices. We consider the pairwise correlation structure between the distances and investigate the large sample properties of the ordinary least square estimator of the model coefficients. Extensive simulations are conducted to evaluate the performance of our method with finite sample size.

Another major component of the thesis is the human microbiome data analysis. We analyze the integrative Human Microbiome Project (iHMP) data set of composition of microbial communities in the digestive tracts of humans by using multiple statistical methods, including our proposed method. The results are presented and interpreted. Existing challenges and future works are also discussed.

ACKNOWLEDGEMENTS

First and foremost I am extremely grateful to my adviser, Dr. Shaoyu Li, whose expertise was invaluable in formulating the research questions and methodology. This thesis would have not been possible without your insightful advice, continuous support, and patience during my PhD study. You went above and beyond to help me continue my research and prepare for job interviews at same time. I cannot express in words how thankful I am.

I would also like to extend my gratitude to Mathematics Department's Graduate Coordinator Dr. Shaozhong Deng, Committee co-Chair Dr. Yanqing Sun, Committee Member Dr. Eliana Christou, and Graduate Faculty Representative Dr. Donna Kazemi for their invaluable feedback and accommodation.

I would also like to thank Mathematics Department's System Administrator Mark Hamrick for his advice as well as providing computational resources. Also, I would like to acknowledge University Research Computing for proving high-performance computing, hpc.

I would also like to extend my appreciation to Graduate School's Assistant Teaching Professor of Writing, Dr. Lisa Russell-Pinson for her notable Dissertation Writing course and Dissertation Writers support group.

Last but not least, I gratefully acknowledge the financial support received towards my PhD from the Graduate School and Mathematics Department.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: DISTANCE–BASED MULTIVARIATE ANALYSIS

Non-standard structured, high dimensional multivariate data are now emerging in many modern research fields, including neuroimaging, ecology, genomics, and human microbiome studies. It is often of great interest to study the functional relationship between these multivariate variables. It could be association between two groups of multivariate variables, such as the relationship between the ecological system and spatial location; or the association between multivariate dependent variable and a univariate independent variable and vice versa, for example the association between the expression level of multiple genes in a molecular functional pathway and a disease outcome. Classical multivariate analysis tools become infeasible because either the massively structured data fail the basic assumptions or how they cluster together between groups of interest in some research field. Powerful statistical tools like multivariate analysis of variance (MANOVA) are based on assumptions of independence, multivariate normal distribution and homogeneity of covariance [1,2]. But many data sets do not conform with these assumptions. Take for example ecological data where number of each species is considered a variable. Abundance of individual species are usually highly aggregated and skewed and non-normally distributed. Also it is common that the number of spices is larger than the sample size (i.e. small n and large p problem) [3,4]. These major challenges lead to two avenue of statistical methodologies for the kind of multivariate data analysis.

For the sake of simplicity and consistency, we will use the following notations through the chapter: let $X_{n\times p}$ be a matrix comprising n observations of p variables. $D_X = (d_{ij}^X)$ denotes the matrix of pairwise distances, where $d_{ij}^X = s_x(X_i, X_j)$ is the pairwise distance/dissimilarity between observation X_i and X_j calculated based on pre-determined distance measurement function $s_x(\cdot)$. Let $Y_{n\times q}$ be a matrix of n observations of the $q-$dimensional dependent variable. $D_Y = (d_{i,j}^Y)$, where $d_{ij}^Y = s_y(Y_i, Y_j)$ is the pairwise distance/dissimilarity between observation Y_i and Y_j determined by a distance function $s_y(\cdot)$.

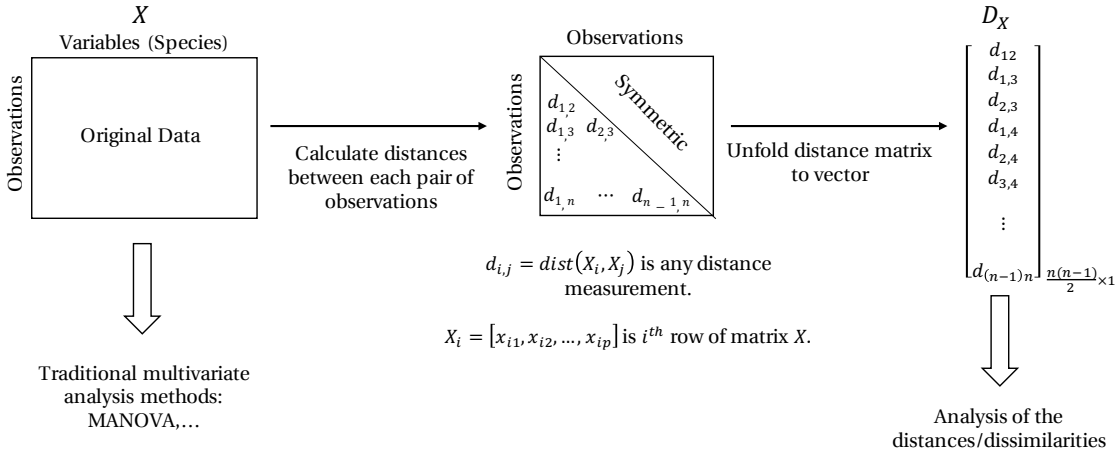


Figure 1.1: Constructing a vector of distances between observations

1.1 Distance based linear correlation

Mantel test Mantel test was primarily introduced [5] to identify time–space clustering of disease. The methods was motivated by a biomedical research problem to identify clustering of leukemia patients in location and time. In the study, X and Y contain observations of the time a location of n patients, respectively. The interpersonal differences in time and location, $D_X = [d_{ij}^X]$ and $D_Y = [d_{ij}^Y]$ are constructed

(Figure 1.1) by using suitable (possibly different) dissimilarity measurements. Because the distance matrices are symmetric, only element in the lower triangle matrix were used to construct the following test statistic:

$$Z(D_X, D_Y) = \sum \sum_{(1 \le i < j \le n)} d_{ij}^X d_{ij}^Y. \tag{1.1}$$

A permutation procedure to mis-match the time and location and obtain a sample of Z values under the null hypothesis of no cluster is proposed.

It is also discussed that the scale of Z will vary from problem to problem; hence, a normalized version of test is:

$$r(D_X, D_Y) = \sum \sum_{1 \le i < j \le n} \frac{(d_{ij}^X - \bar{d}^X)(d_{ij}^Y - \bar{d}^Y)}{var(d^X)^{1/2} var(d^Y)^{1/2}} \tag{1.2}$$

where $d^X = (d_{12}^X, d_{13}^X, \cdots, d_{n-1,n}^X)'$ is the vectorized lower trianle matrix of D^X, and $\bar{d}^X$ denote the its mean value. d^Y and $\bar{d}^Y$ are similarly defined for Y. The resulting statistic r is analogous to the Pearson correlation coefficient between d^X and d^Y [6,7].

A permutation based procedure is used to assess the statistical significance. Specifically, 1. compute the observed value of the statistic $r = r(D_X, D_Y)$ using Equation 1.2. 2. Permute rows and corresponding columns of the distance matrix simultaneously as seen in Figure 1.2 to preserve the symmetry of the structure and construct D_X^*. 3. Compute the Mantel statistic $r^* = r(D_X^*, D_Y)$ for all $n!$ possible permutations or a large random sample of them for large data sets, say $B = 1000$. 4. Calculate a empirical p-value by comparing the observed r value with the B r^* values from permutation, presumably these are samples from the distribution of the test statistic under the null hypothesis that there is no cluster. For a test involving upper-tail, under the null hypothesis of $r \le 0$, the p-value of the test is the proportion of the r^*'s greater than or equal to the observed Mantel statistic r. For a test involving lower-tail, under the null hypothesis of $r \ge 0$, the p-value of the test is the proportion

of the r^*'s less than or equal to the observed Mantel statistic r. For a two-tailed test that is under the null hypothesis of $r = 0$, the p-value of the test is the proportion of the $|r^*|$'s greater than or equal to $|r|$ where $|.|$ is the absolute value function.

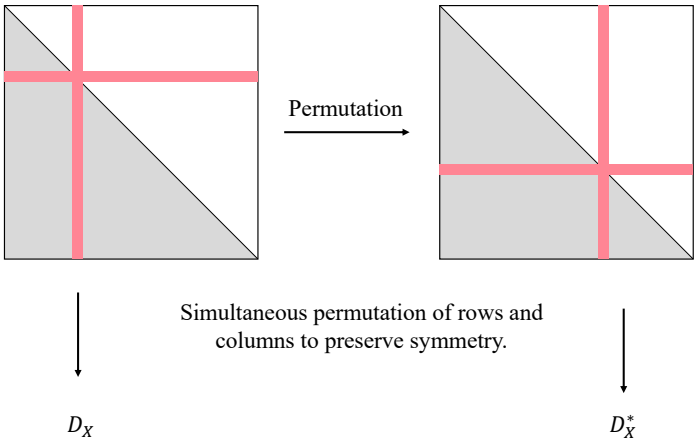


Figure 1.2: Permuting rows and columns jointly to preserve symmetric structure of distance matrix.

Partial Mantel test An extension of Mantel test to three distance matrices that computes partial correlations between the two distance matrices while controlling for the effect of a third distance matrix [7,8]. Given the observed data X, Y and Z, the partial Mantel r statistic is

$$r(D_X, D_Y; D_Z) = \frac{r(D_X, D_Y) - r(D_X, D_Z)r(D_Y, D_Z)}{\sqrt{1 - r(D_X, D_Z)^2}\sqrt{1 - r(D_Y, D_Z)^2}} \tag{1.3}$$

where $r(D_X, D_Y)$ is the simple Mantel statistic calculated by Equation 1.2. Once $r(D_X, D_Y)$, $r(D_X, D_Z)$ and $r(D_Y, D_Z)$ are computed then $r = r(D_X, D_Y; D_Z)$ can be calculated using Equation 1.3. To test hypotheses concerning r, one can perform following steps. Construct D_X^* by random matrix permutation and evaluate $r(D_X^*, D_Y)$ and $r(D_X^*, D_Z)$. Then the partial correlation statistic under permutation, r^*, is the value of $r(D_X^*, D_Y; D_Z)$ computed by Equation 1.3. Repeat the procedure for a large number of times (or possibly for all $n!$ permutations for small data). The empirical p-value of the test is evaluated same way as simple Mantel test.

1.2 Multiple regression on distance matrices

Multiple regression on distance matrices (MRM) An extension of the modeling between multiple pair-wise distance matrices [9]. Besides $Y_{n\times q}$ and $X_{n\times p}$, additional variables can be included, for example, $Z_{n\times k}$. Instead of modeling the linear relationship between the original data, MRM aims to model the relationship between the pairwise distance matrices, D_Y, D_X and D_Z. Because all are symmetric matrices, the lower triangle of each matrix is vectorized to d^X, d^Y, d^Z of the matrices, and a linear regression model is fitted:

$$d_{ij}^{Y} = \beta_0 + \beta_1 d_{ij}^{X} + \beta_2 d_{ij}^{Z} + \varepsilon_{ij}, i = 1, 2, \cdots, n, j = i+1, i+2, \cdots, n. \tag{1.4}$$

Unknown coefficients of the model is estimated by ordinary least square, however, the asymptotic distribution of the t-test for the significance of the parameters for linear regression model on independent observations is not feasible because obviously, the pairwise distances are correlated. A permutation based procedure is also suggested to assess the significance of the test: The dependent distance matrix is permuted while holding the explanatory distance matrices unchanged. Let $\hat{\beta}_i^*$ be the estimated regression coefficients in a permutation. Under the null hypothesis $\beta_i = 0$, p-value is the proportion of $\hat{\beta}_i^*$'s larger in absolute value than observed $\hat{\beta}_i$, for all possible or a large sample of permutations.

1.3 Distance based multivariate analysis of variance

Generalizations of MANOVA Powerful multivariate statistical methods, such as

the classical multivariate analysis of variance (MVANOVA), have existed for decades [10–15]. However, in some applications, the data structure fails the fundamental assumptions of these methods. For example, abundance data in ecological studies, take discrete values, rather than being continuous, rare species contribute lots of zeros to the data set, more variables than sampling units. These kind of multivariate response variable fails the traditional MANOVA because 1) the dimension of the variable p is generally greater than sample size n, so the sample variance covariance matrix becomes singular, so traditional tools, including Hottelling's T^2 [10] and Wilki's Lambda test [11] can not be used.

There are two ways to resolve the issue, either take the generalized inverse or go with Dempster trace criterion [16, 17] for one and two sample cases. For inference, the F-approximation [18]. Gower [19, 20], Gower and Legendre [21], and Gower and Krzanowski [22] investigated the connection between sample variance and distance, specifically, Euclidean distance, $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i<j}^{n}(x_i - x_j)^2$. They proposed to extend all the necessary features of MANOVA to the dissimilarity matrix, for example, decompose the total variance to be between-group and within-group by using the dissimilarity matrix. The idea has been carried on and motivates various following works on pairwise distance matrix, among them, the PERMANOVA [4] gains the most attention due to its use in recent microbiome studies.

Permutational MANOVA Permutational Multivariate Analysis of Variance (PERMANOVA) [4] is a non–parametric analogue to traditional MANOVA statistic obtained by partitioning the lower triangle of a distance matrix into between-group distances and within-group distances (Figure 1.3).

Let $N = an$ be the total number of observations, where a is the number of groups, and n is the number of observations in each group. Let d_{ij} be the distance between

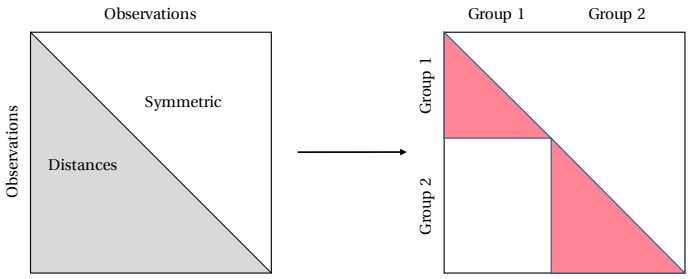


Figure 1.3: Partitioning a distance matrix into groups. In the right panel, red triangles represent within-group and the white rectangle represents the between-group.

sample units i and j. The total sum of squares is defined as

$$SS_T = \frac{1}{N} \sum \sum_{(1 \leq i < j \leq N)} d_{ij}^2$$

and within-group or residual sum of squares is

$$SS_W = \frac{1}{n} \sum \sum_{(1 \leq i < j \leq N)} d_{ij}^2 I_{ij},$$

where $I_{ij} = 1$ if observations i and j are in same group, and $I_{ij} = 0$, otherwise. Then, the between-group sum of squares is defined as

$$SS_A = SS_T - SS_W.$$

A pseudo F-ratio test statistic is computed by

$$F = \frac{SS_A/(a-1)}{SS_W/(N-a)}$$

Anderson and Walsh [23] state PERMANOVA assumes exchangeabililty, that is joint conditional distribution $p(X_1, X_2, \cdots, X_n | Y_i = y_i, \textit{ for all } i)$ is invariant under permutation of the sample units among the groups [24]. The null hypothesis tested by PERMANOVA is H_0: centroids of groups as defined in space of chosen resemblance

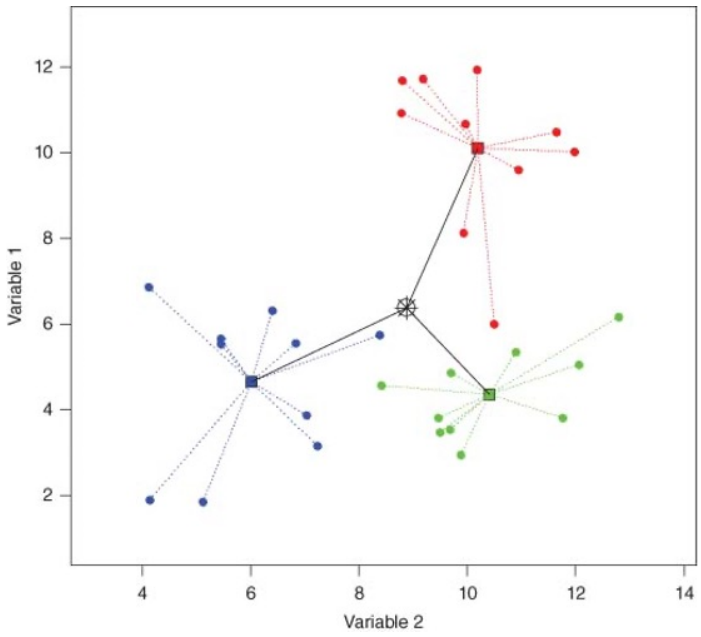


Figure 1.4: Schematic diagram of geometric partitioning for PERMANOVA, shown for g=3 groups of n=10 sampling units per group in two-dimensional (bivariate, p= 2) Euclidean space. First published: `https://doi.org/10.1002/9781118445112.stat07841`

measure are equivalent for all groups. They later explain that, if H_0 were true, centroids of each group is within same distance to the overall centroid (Figure 1.4) [23, 25]. Empirical p-value is obtained via permutation. For each permutation of the data the permutation test statistic F^{π} is calculated and p-value is computed using

$$p - value = \frac{\#(F^{\pi} \geq F) + 1}{\#(F^{\pi}) + 1}.$$

1.4 Limitations of currently published methods

Variations of Mantel test are usually used in spatial correlation analysis, for example the study of geographical genetic divergence in botanical science. Though, Mantel

test suffers lack of power for spatially autocorrelated data [26]. Spatial autocorrelation is the term used to describe the presence of systematic spatial variation in a variable. Positive spatial autocorrelation is the tendency for areas or sites that are close together to have similar values [27]. Inflated Type-I error and relatively low power appears to be a general feature of the Mantel test [28–30].

PERMANOVA is routinely used in numerical ecology to test for the location differences in microbial communities. It has been discussed PERMANOVA fails to detect a significant between-group difference unless it is present in taxa (units of any rank i.e. kingdom, phylum, class, order, family, genus and species, designating an organism or a group of organisms) with high variance [31]. Moreover, PERMANOVA is prone to inflated Type-I error in presence of heteroscedasticity [32, 33].

Franckowiak *et al.* in [34] examined the ability of model selection criteria based on Akaike's information criterion (AIC), its small-sample correction (AICc), and the Bayesian information criterion (BIC) to reliably rank candidate models when applied with MRM while varying the sample size and strongly discourage the continued application of AIC, AICc, and BIC for model selection with MRM.

A possible explanation of challenges summarized above could be misusing naive permutation testing. The concept of permutation relies on exchangeability and simple permutation could lead to inflated Type-I error or low power [35, 36], particularly, in presence of unequal variances, correlations, skewness, or unequal sample sizes in two groups [37, 38].

CHAPTER 2: HUMAN MICROBIOME AND INFLAMMATORY BOWEL DISEASES

2.1 Introduction

"What is the human microbiome?" has troubled researchers [39] since Lederberg's coinage of "microbiome" in 2001 [40]. Researchers have confused how to exactly define the human microbiome and interchangeably used terminologies "microbiota" and "microbiome". The term microbiota is referred to the microbial taxa associated with humans to signify the communities of microorganisms within a specific environment [41]. The term microbiome is defined as the collection of the microbial taxa or microbes and their genes [39], [42], the entire microbial communities. Depending on the collection of microbes, i.e. the body site in which those microorganisms inhabit in, researchers use specific terms such as gut microbiome, gut microbiome or oral microbiome.

Most sources cite the number of human cells as 10^{13} or 10^{14}, and a recent study reported 3.7×10^{13} human cells in a reference human [43]. Estimates for the number of microbial cells in the body are usually $10^{14} - 10^{15}$ [44], [45] which suggests a ratio of $10 : 1$ microbial cells to human cells. However, more recent studies suggest the ratio to be much closer to $1 : 1$ [46].

Inflammatory Bowel Disease (IBD) is a broad term that describes conditions char-

acterized by chronic inflammation of the gastrointestinal tract. The two most common inflammatory bowel diseases are ulcerative colitis (UC) and Crohn's disease (CD). Inflammation affects the entire digestive tract in Crohn's disease and only the large intestine (also called the colon) in ulcerative colitis. In 2015, more than 3.5 million people worldwide were diagnosed with IBD (either Crohn's disease or ulcerative colitis) [47]. Growing evidence suggests that gut microbiota may be an important factor in the pathogenesis of a variety of diseases including inflammatory [48]. Dysbiosis of the gut microbiota, an alteration of the microbial community structure associated with disease, has been consistently observed in patients with IBD. Although the dysbiosis may simply be a result of the inflammatory process [32], it may play a role in the pathogenesis of disease where there is an increase in potentially harmful bacterial and a reduction in more protective bacterial species [49].

A paragraph about alpha diversity and beta diversity for microbiome data:

2.2 The human microbiome project

The National Institutes of Health (NIH) Human Microbiome Project (HMP, `https://hmpdacc.org`) has been carried out over ten years and two phases to provide resources, methods, and discoveries that link interactions between humans and their microbiomes to health–related outcomes. The second phase of the HMP, the Integrative HMP (iHMP or HMP2) [50], was designed to explore host-microbiome interplay, including immunity, metabolism, and dynamic molecular activity, to gain a more holistic view of host-microbe interactions over time [51]. The iHMP projects included three studies that followed the dynamics of human health and disease during conditions with known microbiome interactions. These comprised pregnancy and preterm birth (PTB) [52]; inflammatory bowel diseases (IBD) [53]; and stressors that affect individuals with prediabetes [54].

A collection of commentary and research publications from across Nature journals and related publications from HMP2 can be found at `https://www.nature.`

com/collections/fiabfcjbfj, and a rich multi-omic data resource at http://www.ihmpdcc.org.

2.3 HMP2 Study: gut microbiome and inflammatory bowel diseases

The Inflammatory Bowel Disease Multi'omics Database (IBDMDB) project followed 130 (Table 2.1) individuals from five clinical centres over the course of one year each as part of HMP2. Integrated longitudinal molecular profiles of microbial and host activity were generated by analysing 1,785 stool samples (self-collected and sent by mail every two weeks), 651 intestinal biopsies (collected colonoscopically at baseline), and 529 quarterly blood samples [51]. Multiple molecular profiles were generated from the same sets of samples, including stool metagenomes, metatranscriptomes, metaproteomes, viromes, metabolomes, host exomes, epigenomes, transcriptomes, and serological profiles, among others, allowing concurrent changes to be observed in multiple types of host and microbial molecular and clinical activity over time. Protocols and results from the study, further information about its infrastructure, and both raw and processed data products are available through the IBDMDB data portal http://ibdmdb.org, from the HMP2 Data Coordination Center (DCC; http://ihmpdcc.org). [53, 55–59]

Table 2.1: Number of Subjects by Category in HMP2 study of inflammatory bowel diseases (IBD). nonIBD: not diagnosed with IBD or control group; CD: diagnosed with Crohn's disease; UC: diagnosed with ulcerative colitis.

Disease	Gender	Child	School-age	Adult	Senior	Total
nonIBD	female	3	5	4	0	12
	male	3	2	9	1	15
CD	female	3	8	20	1	32
	male	5	15	13	0	33
UC	female	2	3	14	1	21
	male	1	9	7	1	17
Total		17	42	67	4	130

2.4 Metagenomes: profiling and statistical analyses

National Human Genome Research Institute (NHGRI, `https://www.genome.gov/genetics-glossary/Metagenomics`) defines metagenomics as the study of a collection of genetic material (genomes) from a mixed community of organisms. Metagenomics usually refers to the study of microbial communities. The genome is the entire set of genetic instructions found in a cell. Genomics is obtaining the DNA sequence, and meta implies of many organisms together. Metagenomics is usually used in the study of microbial communities where one can't separate one microbe from another.

As a part of HMP2, the composition of microbial communities of stool samples were profiled from metagenomic shotgun sequencing data and MetaPhlAn (Metagenomic Phylogenetic Analysis) [60]. Processing microbiome data generates a matrix that relates feature abundance (taxa or genes) to samples. The microbiome data are highly dimensional, often representing thousands of different taxa, and sparse and zeros inflated matrix [61]. Stool samples were collected over the course of one year metagenomic profiles were generated and classified at seven taxonomic ranks—the relative level of a group of organisms (a taxon) in a taxonomic hierarchy, species, genus, family, order, class, phylum and kingdom. Bacteria make up for over 99% of detected microorganisms. Table 2.2 shows the prevalence (number of samples representing the taxon) of the phyla of gut bacteria. Firmicutes and bacteroidetes are present in almost all samples whereas spirochaetes, chlamydiae or cyanobacteria appear in less than 1% of samples.

Figure 2.1 shows the relative abundance of bacteria at phylum level for adults in three groups, diagnosed with Crohn's disease (CD), ulcerative colitis (UC) or healthy controls (nonIBD). On average, bacteroidetes and firmicutes make up for 95% of phylum level of gut bacteria.

Table 2.2: Abundances and relative abundances of the phyla of gut bacteria

	Abundance	Relative abundance
Firmicutes	1602	1.00
Bacteroidetes	1572	0.98
Proteobacteria	1499	0.93
Actinobacteria	1440	0.90
Verrucomicrobia	502	0.31
Fusobacteria	76	0.05
Synergistetes	45	0.03
Lentisphaerae	17	0.01
Candidatus Melainabacteria	12	0.01
Spirochaetes	5	0.00
Chlamydiae	2	0.00
Cyanobacteria	1	0.00

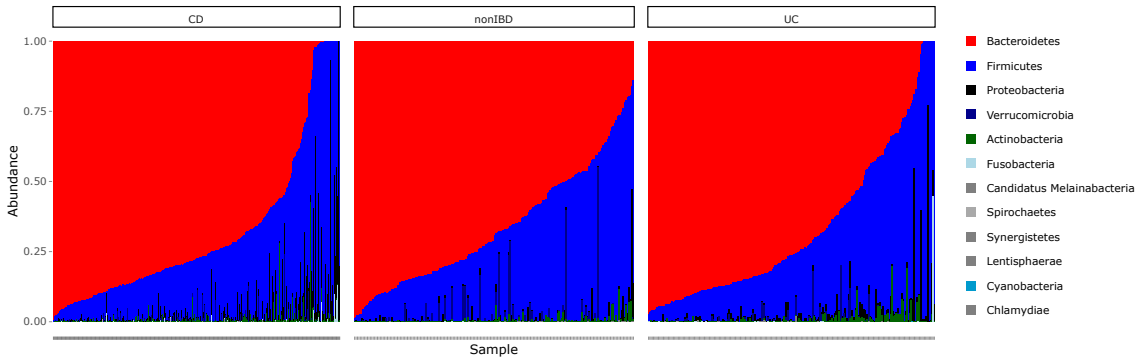


Figure 2.1: Abundance of phyla of gut bacteria for adults with CD ($n = 355$), nonIBD ($n = 196$) and UC ($n = 231$)

Over 580 distinct species are represented in this data. Attributes such as species richness, evenness and diversity can be used to compare community compositions. Species richness is the number of different species community. Species evenness is a description of the distribution of abundance across the species in a community. Species diversity is usually described by an index that includes both richness and evenness of the species. Global taxonomic richness is affected by variation in three components: within-community, or alpha diversity, between-community, or beta diversity, and between-region, or gamma diversity [62–69].

Shannon's diversity index is commonly used in ecology as a measure of alpha diver-

sity. It's based on the Shannon's entropy formula, $H = -\sum_{i=1}^{S} p_i \ln p_i$ where $p_i = \frac{n}{N}$ is the proportion of the number of individual species i found (n) divided by the total number of individuals found (N) and S is the number of different species. Figure 2.2 shows the distribution of Shannon indices for each cohort. Kruskal-Wallis [70] rank sum test of differences in mean values of Shannon indices as shown on the plot ($\chi_2^2 = 41.078$, p-value ≈ 0).

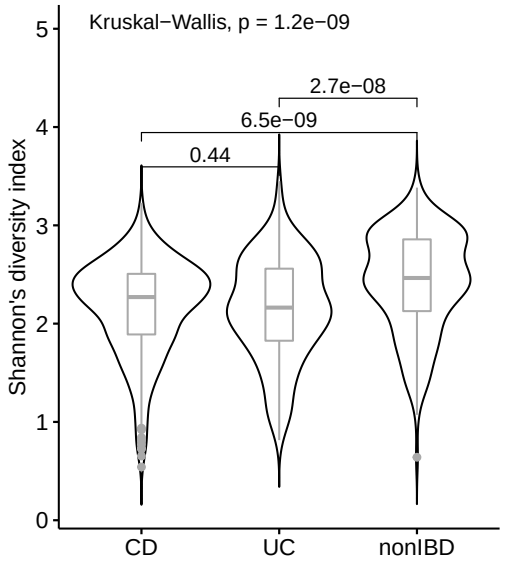


Figure 2.2: Distribution of Shannon's diversity indices

Common indices of beta diversity include Bray-Curtis, Unifrac distance. Bray-Curtis distance [71], $BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$, is a commonly used distance measure in ecological data, where S_i is the total number of specimens counted on site i, S_j is the total number of specimens counted on site j, and, C_{ij} is the sum of only the lesser counts for each specimen found on both sites. Weighted/unweighted UniFrac [72–75] is method for computing differences between microbial communities based on phylogenetic information. UniFrac, measures the phylogenetic distance between sets of taxa in a phylogenetic tree as the fraction of the branch length of the tree that leads to descendants from either one environment or the other, but not both. Weighted UniFrac accounts for abundance of observed organisms whereas unweighted

UniFrac only considers their presence or absence. A schematic diagram of UniFrac calculations is shown in Figure 2.3

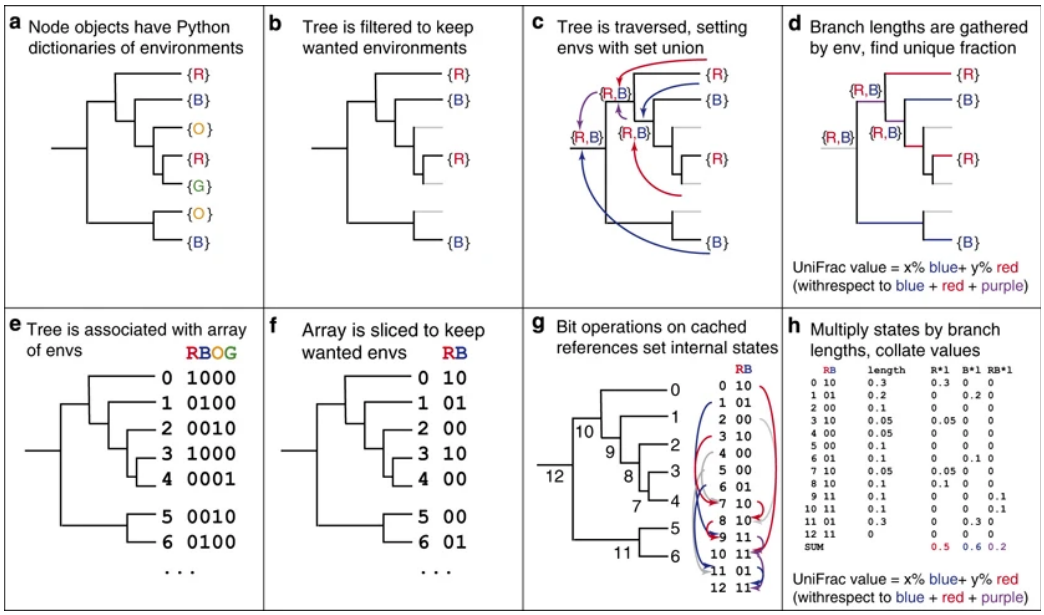
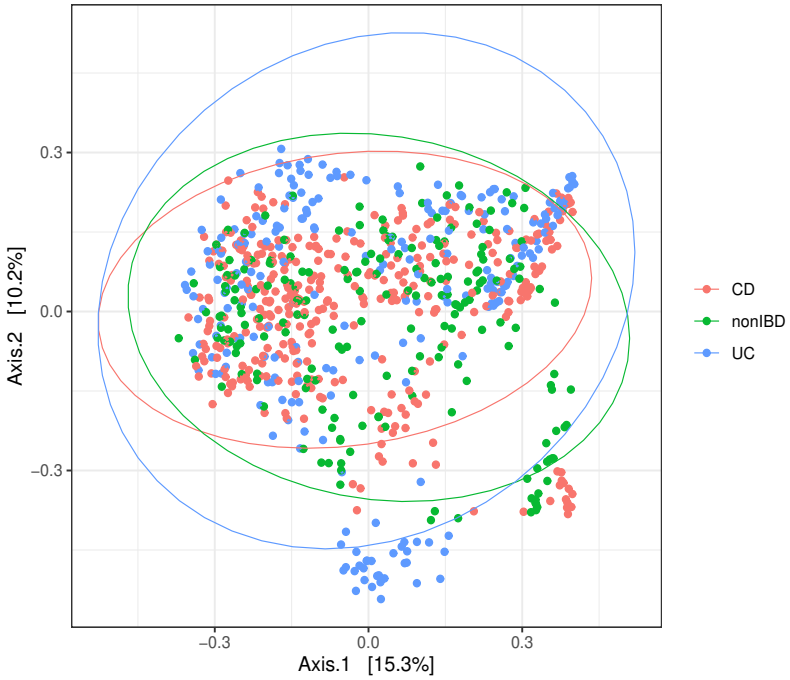


Figure 2.3: Schematic diagram of UniFrac calculations. First published: `https://doi.org/10.1038/ismej.2009.97`

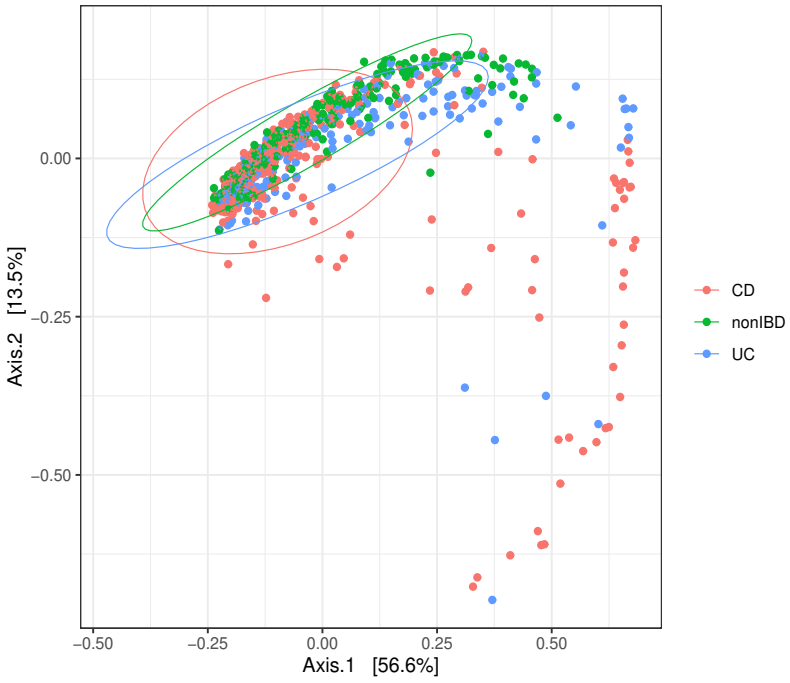
A well-known tool to visualize high dimensional community ecology data is principal coordinates analysis. Principal coordinates analysis (PCoA) is a metric multidimensional scaling method based on projection, which uses spectral decomposition to approximate a matrix of distances/dissimilarities by the distances between a set of points in few dimensions [76]. PCoA is equivalent to principal component analysis (PCA) when euclidean distances are used. Plots of first two axes of PCoA accompanied by 95% student's-t confidence ellipses were generated and displayed Figure 2.4 using (a) Bray-Curtis and (b) Weighted UniFrac distance methods. Different colors indicate the cohort in which the sample lies.

Figure 2.4: PCoA visualization of samples-wise (a) Bray-Curtis and (b) WUniFrac distances color coded for disease type accompanied by 95% student's-t confidence ellipses.

(a) Bray-Curtis



(b) WUniFrac



Weighted Unifrac distance seems to be separate cases from control with 70% variation explained by first to axes in contrast to Bray-Curtis distance with only 25.5% variation explained by first to axes and all samples overlapping.

A PERMANOVA was performed using three group: CD and UC cases and nonIBD controls using Bray-Curtis distance summarized in Table 2.3 and Weighted UniFrac distance summarized in Table 2.4.

Table 2.3: PERMANOVA using B-C based on 9999 permutations; terms added sequentially (first to last)

	Df	Sums of Sqs	Mean Sqs	F.Model	Pr(>F)
CD	1	1.981	1.9808	6.5911	0.0001
UC	1	3.822	3.8223	12.7185	0.0001
Residuals	779	234.114	0.3005		
Total	781	239.918			

Table 2.4: PERMANOVA using WUnifrac based on 9999 permutations; terms added sequentially (first to last)

	Df	Sums of Sqs	Mean Sqs	F.Model	Pr(>F)
CD	1	0.742	0.74246	8.3305	0.0001
UC	1	0.580	0.57972	6.5045	0.002
Residuals	779	69.429	0.08913		
Total	781	70.751			

Intuitively, box plots of relative abundances and distance measurements are presented in Figure 2.5 and Figure 2.6. Since there were over 500 distinct species with high sparsity, box plots were generated at phylum level for a better resolution. Figure 2.5 displays the relative abundance of phyla grouped by UC and CD cases and nonIBD controls within adults participated in HMP2 study. Bacteroidetes and firmicutes, proteobacteria, actinobacteria and verrucomicrobia have highest relative abundances (core taxa) and therefore of interest to be investigated. Figure 2.6 displays the distances/dissimilarities of each phylum labeled by between and within cohorts. Bacteroidetes and firmicutes distance measurements present higher interquartile range and proteobacteria, actinobacteria and verrucomicrobia distances are heavily skewed.

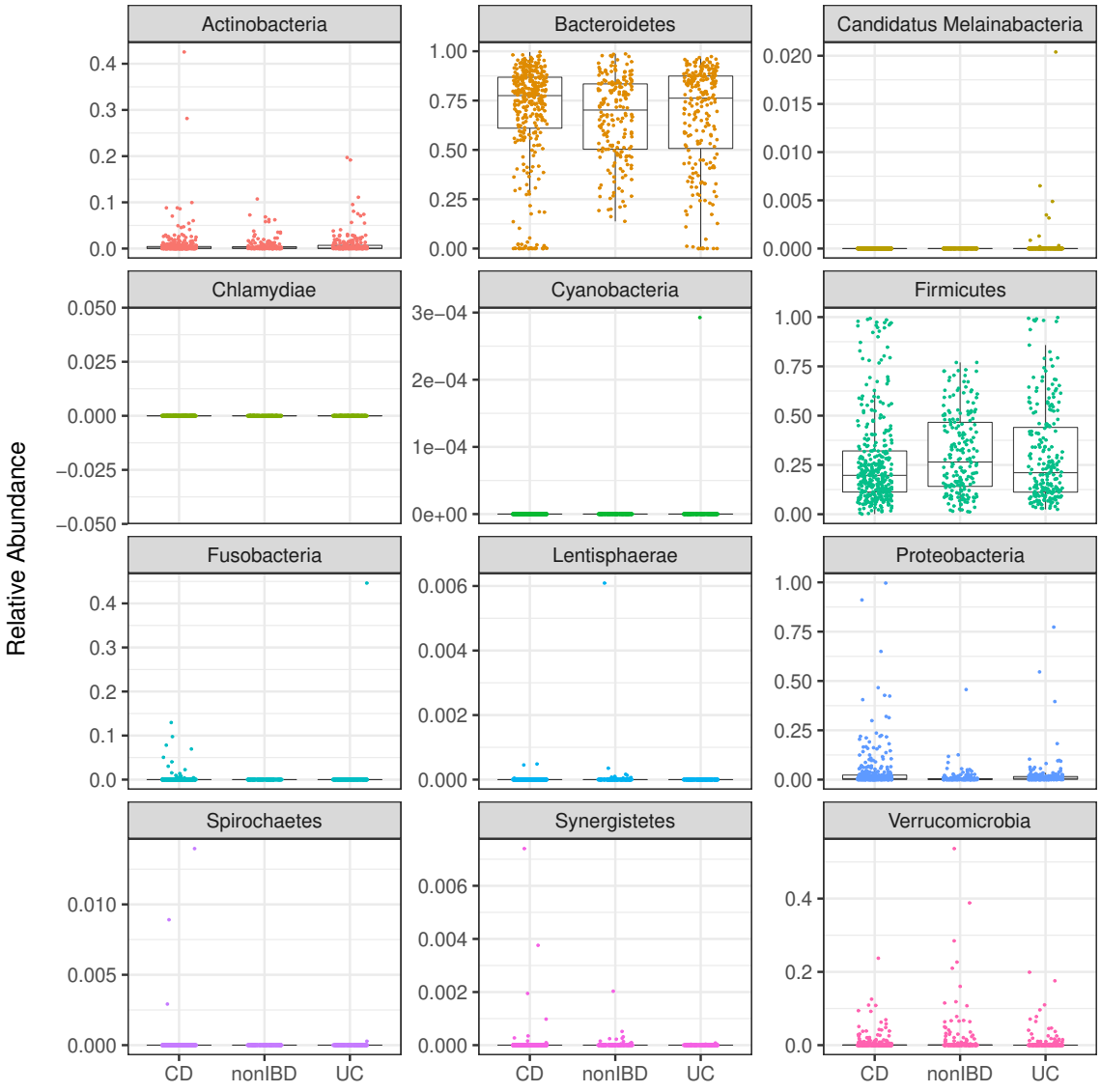


Figure 2.5: Box plots of relative abundance of gut microbiome of adults at phylum level for UC and CD cases and nonIBD controls.

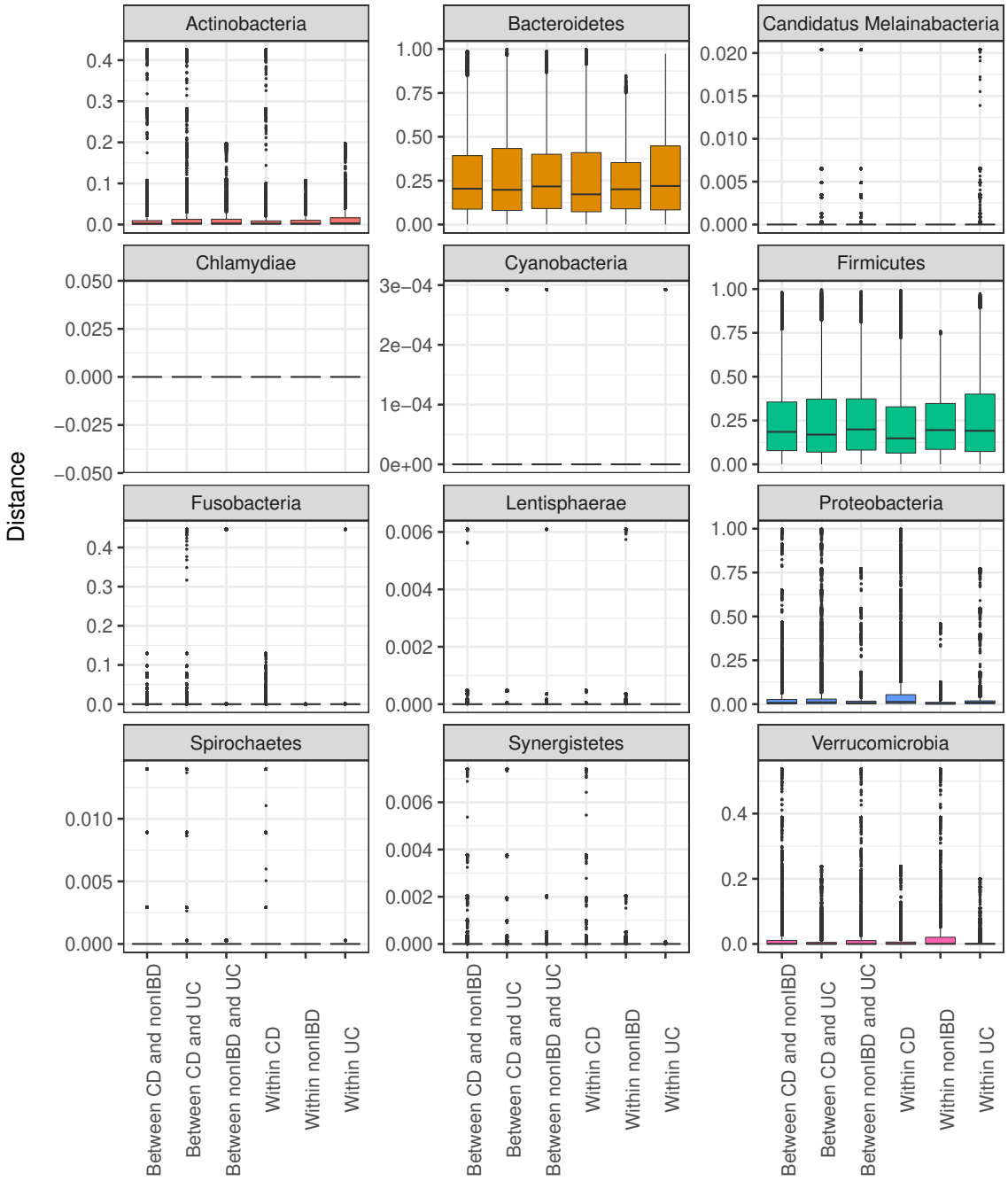


Figure 2.6: Box plots lay out the distribution of pairwise distances of gut microbiome of adults at phylum level, labeled by between and within cohorts.

Mantel test (Table 2.5) was used to investigate whether lower dissimilarities of species with phylum level bacteroidetes and firmicutes correspond to lower(or higher) dissimilarities of species with phylum level proteobacteria, actinobacteria and ver-

rucomicrobia for each group. The null hypothesis for "pval1" is that the Mantel r statistic will be equal to or smaller than zero, i.e. negative correlation. Conversely, the null hypothesis for "pval2" is that the Mantel r statistic is equal to or greater than zero, i.e. positive correlation. The null hypothesis for "pval3" is that the Mantel r statistic is equal to zero, i.e. no correlation.

Table 2.5: Mantel r-statistic and p-values: "pval1" under $H_0 : r \leq 0$; "pval2" under $H_a : r \geq 0$; "pval3" when $H_0 : r = 0$; accompanied by lower and upper limits of 95% CI

	Mantel r	pval1	pval2	pval3	llim.2.5%	ulim.97.5%
within CD	0.0423	0.0086	0.9915	0.0106	0.0282	0.0572
within UC	0.0688	0.0008	0.9993	0.0008	0.0479	0.0863
within nonIBD	0.0991	0.0001	1.0000	0.0001	0.0851	0.1128

We learn that dissimilarities of species with phylum level bacteroidetes and firmicutes are positively correlated to dissimilarities of species with phylum level proteobacteria, actinobacteria and verrucomicrobia for each cohort. It is of interest to study the biological relationships and interactions of species grouped as above. Smaller Mantel r for UC and CD cases might be a bilogical signal that needs further investigations. For application of MRM on this data set, distances at species level were computed and labeled as withing and between UC, CD and nonIBD groups. Considering the within nonIBD distances as reference group, the MRM model is

$$\begin{aligned} d_{ij} = \beta_0 &+ \beta_1 I_{\text{Within CD}} \\ &+ \beta_2 I_{\text{Between CD and UC}} \\ &+ \beta_3 I_{\text{Between CD and nonIBD}} \\ &+ \beta_4 I_{\text{ Within UC}} \\ &+ \beta_5 I_{\text{Between nonIBD and UC}} + \varepsilon_{ij} \end{aligned} \tag{2.1}$$

where $I(.)$ is the indicator function. MRM coefficients and p-values based on 9999 permutations are summarized for choice of distance Bray-Curtis in Table 2.6 and for Weighted UniFrac in **??**.

Table 2.6: MRM coefficients and p-values with 9999 permutations when B-C distance is used.

	Distance	pval
Int	0.7061	1.0000
Within.CD	0.0622	0.0001
Between.CD.and.UC	0.0742	0.0001
Between.CD.and.nonIBD	0.0552	0.0001
Within.UC	0.0699	0.0002
Between.nonIBD.and.UC	0.0632	0.0001

Table 2.7: MRM coefficients and p-values with 9999 permutations when WUnifrac distance is used.

	Distance	pval
Int	0.3174	0.9972
Within.CD	0.0525	0.0295
Between.CD.and.UC	0.0571	0.0118
Between.CD.and.nonIBD	0.0390	0.0017
Within.UC	0.0531	0.0460
Between.nonIBD.and.UC	0.0367	0.0076

Difference in p-value obtained on different distances indicates that MRM may suffer inflated Type-I error rate when Bray-Curtis distance is used.

CHAPTER 3: A DISTANCE-BASED LINEAR REGRESSION MODEL

3.1 Introduction

3.1.1 Statistical model and parameter estimation

Suppose that we have n independent data draws denoted as $(\mathbf{x}_i, \mathbf{y}_i), i = 1, 2, \cdots, n$, where $\mathbf{x}_i \in \mathcal{R}^p$ and $\mathbf{y}_i \in \mathcal{R}^q$. Pairwise distances or dissimilarities between all combinations of n objects are constructed and denoted as $\mathbf{D}_{\mathbf{Y}} = (y_{ij})_{n\times n}, \mathbf{D}_{\mathbf{x}^{(1)}} = (x_{1,ij}), \cdots, \mathbf{D}_{\mathbf{x}^{(K)}} = (x_{K,ij})$, respectively. Where $K \leq p$ refers to the number of subgroups of $\mathbf{x}_i$, that is, $\mathbf{x}_i = (\mathbf{x}_i^{(1)}, \cdots, \mathbf{x}_i^{(K)})'$. Depending on the application, the pairwise distance measurements may reflect distance in species abundances, geographical location, and genetic distance using compound or individual distance measure. The pairwise distance transforms the original independent observations $(\mathbf{x}_i, \mathbf{y}_i), i = 1, 2, \cdots, n$ to correlated pairwise distances, which are denoted as $y_{ij} = s_y(\mathbf{y}_i, \mathbf{y}_j)$ and $x_{k,ij} = s_k(\mathbf{x}_i^{(k)}, \mathbf{x}_j^{(k)}), k = 1, 2, \cdots, K$. Assuming the distances are symmetric, that is, $d_{ij} = d_{ji}$, the upper triangle of pairwise distance matrices are unfolded and vectorized. We then model the relationship between y_{ij} and $x_{k,ij}$'s via the following regression model:

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \varepsilon_{ij} \quad 1 \leq i < j \leq n, \tag{3.1}$$

here, $\mathbf{x}_{ij} = (1, x_{1,ij}, \cdots, x_{K,ij})'$ and $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_1, \beta_K)'$, is the vector of coefficients.

A least square estimator of the regression coefficients $\boldsymbol{\beta}$ can be derived as in classic linear regression model for independent observations, by minimizing the following sum of squares:

$$U_n(\boldsymbol{\beta}) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} (y_{ij} - \mathbf{x}'_{ij}\boldsymbol{\beta})^2 = \frac{1}{\binom{n}{2}} (Y - X\boldsymbol{\beta})'(Y - X\boldsymbol{\beta}), \tag{3.2}$$

where

$$Y = (y_{12}, y_{13}, \cdots, y_{1n}, y_{23}, \cdots, y_{n-1,n})'$$

and

$$X = (\mathbf{x}_{12}, \mathbf{x}_{13}, \cdots, \mathbf{x}_{1n}, \mathbf{x}_{23}, \cdots, \mathbf{x}_{n-1,n})'.$$

The ordinary least square estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y$. This is the multivariate regression model (MRM) of pairwise distances [9] that we reviewed in Chapter 1. The model was motivated by ecological studies and is more flexible in terms incorporating multiple distance matrices. Not only different types of data, such as counts, binary, continuous and categorical, that may be analyze, it also allows the covariates to be separated into sub-groups, which may be desired in real applications and include the multiple distance matrices of the groups of the covariates in the model. For example, separating geographic distance from other environmental variables when studying their effects on similarity between ecological communities.

Although the parameter estimation is straightforward and, the large sample properties of the estimator is different from the classical linear regression model because y_{ij} are apparently correlated. Lichstein [9] applied a permutation based procedure to test the significance of every individual model coefficient. However, a well known pitfall of permutation method is that it is computationally expensive because the model needs to be re-fitted for a large number of times. Especially in large scale studies, such as microbiome genome-wide association studies, the association between microbiome

community and tens of thousands of genetic markers are tested. And multiple testing corrections are always required, which means an empirical p-value smaller than 10^{-6} can be called significant. In that case, at least 10^6 permutations are needed for each test .

Therefore, we investigate the large sample properties, including asymptotic consistency and normality, of the least square estimator of MRM in the following section. Based on the derived theoretical results, a computationally much efficient inference procedure is developed.

3.1.2 Large sample theory

Observing that the sum of squares of the linear regression model for pairwise distance matrices has the form of a second-order U-statistics, we are motivated to use the large sample properties of U-statistics [77–80] to study the theoretical properties of $\hat{\boldsymbol{\beta}}$.

Theorem 1 (Asymptotic consistency)**.** *Assuming Θ is compact, then the least square estimator defined by minimizing $U_n(\boldsymbol{\beta})$, $\hat{\boldsymbol{\beta}}$, is consistent for $\boldsymbol{\beta}_0 = argmin_{\boldsymbol{\beta}} E(U_n(\boldsymbol{\beta}))$.*

Proof. Since $U_n(\boldsymbol{\beta})$ is a second order U-statistic, by the strong law of large numbers for U-statistic, $U_n(\boldsymbol{\beta}) \to E(U_n(\boldsymbol{\beta}))$ almost surely. And it is also trivial to prove that $\boldsymbol{\beta}_0$ is the unique minimizer of $U(\boldsymbol{\beta}) = E(U_n(\boldsymbol{\beta}))$. Then the consistency can be derived by following the argument for consistency of M-estimators. □

Theorem 2 (Asymptotic Normality)**.** *If $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}_0$, then $\hat{\boldsymbol{\beta}}$ is asymptotic normal*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to N(\mathbf{0}, \Sigma) \tag{3.3}$$

where

$$\Sigma = 4H_0^{-1}V_0H_0^{-1}$$

with

$$H_0^{-1} = E(\mathbf{x}_{ij}\mathbf{x}_{ij}')$$

and

$$V_0 = Var(E[\mathbf{x_{ij}}(y_{ij} - \mathbf{x'_{ij}}\boldsymbol{\beta}_0)|\mathbf{x_i}, y_i]).$$

Proof. Define a normalized score function:

$$Q_n(\boldsymbol{\beta}) = \sqrt{n}\frac{\partial U_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{\sqrt{n}}{\binom{n}{2}} \sum_{1\le i<j\le n} \mathbf{x}_{ij}(y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta}) \tag{3.4}$$

Taylor series expansion of $Q_n(\hat{\boldsymbol{\beta}})$ at $\boldsymbol{\beta}_0$:

$$Q_n(\hat{\boldsymbol{\beta}}) = Q_n(\boldsymbol{\beta}_0) + \frac{\partial Q_n(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_p(1) \tag{3.5}$$

therefore,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = -H_n^{-1}Q_n(\boldsymbol{\beta}_0) + o_p(1) \tag{3.6}$$

where $H_n = \frac{1}{\binom{n}{2}}X'X$.
According to the theories of second order U-statistics [81],

$$Q_n(\boldsymbol{\beta}_0) = \hat{Q}_n(\boldsymbol{\beta}_0) + o_p(1) \tag{3.7}$$

where

$$\begin{aligned}
\hat{Q}_n(\boldsymbol{\beta}_0) &= \sum_{i=1}^{n} E(Q_n(\boldsymbol{\beta}_0)|y_i, \mathbf{x}_i) \\
&= -\sum_{i=1}^{n} \sqrt{n}\binom{n}{2}^{-1} \sum_{1\le i<j\le n} E[\mathbf{x}_{ij}(y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta}_0)|y_i, \mathbf{x}_i] \\
&= -\sqrt{n}\binom{n}{2}^{-1} \sum_{i=1}^{n} (\binom{n-1}{1} E[\mathbf{x}_{ij}(y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta}_0)|y_i, \mathbf{x}_i]) \\
&= -\frac{2}{\sqrt{n}} \sum_{i=1}^{n} r(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0)
\end{aligned} \tag{3.8}$$

Therefore,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = -H_n^{-1}\frac{2}{\sqrt{n}}\sum_{i=1}^{n} r(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0) + o_p(1) \tag{3.9}$$

Since $r(y_i, \mathbf{x}_i)$ are i.i.d, by the central limit theorem,

$$\frac{2}{\sqrt{n}}\sum_{i=1}^{n} r(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0) \to N(\mathbf{0}, 4V_0) \tag{3.10}$$

where $V_0 = var(r(y_i, \mathbf{x}_i))$.

By the law of large of large number, $H_n \to E(\mathbf{x}_{ij}\mathbf{x}'_{ij}) \equiv H_0$. The Slutsky theorem implies that

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to N(\mathbf{0}, 4H_0^{-1}V_0H_0^{-1}). \tag{3.11}$$

□

3.1.3 Estimating the covariance matrix and hypothesis testing

We estimate the conditional expectation $r(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0)$ by

$$\widehat{r(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0)} \equiv \binom{n-1}{1}^{-1} \sum_{j \neq i} \mathbf{x}_{ij}(y_{ij} - \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}}) \tag{3.12}$$

then,

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{r(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0)} = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \mathbf{x}_{ij}(y_{ij} - \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}}) = -Q_n(\hat{\boldsymbol{\beta}})/\sqrt{n} = 0. \tag{3.13}$$

And

$$\hat{V}_n \equiv \frac{1}{n}\sum_{i=1}^{n} \hat{r}(y_i, x_i, \hat{\boldsymbol{\beta}})\hat{r}(y_i, x_i, \hat{\boldsymbol{\beta}})' \tag{3.14}$$

is an unbiased consistent estimator of V_0 [81].

Hypothesis testing: Hypothesis testing is an essential component of statistical inference because it is often of practical interest to test if a certain covariate is significantly associated with the response variable. This testing problem can be ac-

commodated by considering the form: $H_0 : \beta_k = 0$ vs $H_1 : \beta_k \neq 0$, if the coefficient of the k^{th} covariate in the regression model is zero. Since we have derived the asymptotic normality of the least square estimator of $\boldsymbol{\beta}$, we may use a Wald type test statistic $T_k = \frac{\hat{\beta}_k}{se(\hat{\beta}_k)}$. P-value of the test can be calculated by using the normal approximation based on the asymptotic normality result.

3.2 Simulation studies

In this section, we aimed to assess the accuracy of estimation as well as investigate the finite sample performance of our proposed testing procedure by conducting extensive simulation studies. We considered six different scenarios to evaluate the performance of our method from multiple perspectives. Simulation set-up and results of each scenario are detailed in the following sub-sections.

Simulation for Assessing the Accuracy of Estimation

Scenario I: In this simulation, we mimic genetic association studies and generate genotype data of 20 single nucleotide polymorphisms (SNPs). For each SNP, $\mathbf{x}_i \sim multinomial(p = (maf_i^2, 2maf_i(1-maf_i), (1-maf_i)^2))$, as detailed in Figure 3.1. Here, maf_i is short for minor allele frequency of the i^{th}, which is the frequency at which the rare allele occurs in diploid human genome. Three different values of mafs are considered ($maf = 0.1, 0.3, 0.5$) in this simulation. We calculate distance genomic distance between individual i and j by $x_{ij} = \sum_{k=1}^{20} |x_{ik} - x_{jk}|$, and simulate observations of the response variable via the following model: $y_{ij} = 0.2x_{ij} + \varepsilon_{ij}$ where $\varepsilon_{ij} = |\varepsilon_i - \varepsilon_j|$ for each $\varepsilon_i \sim N(0, .01^2)$.

Figure 3.1: An intuitive illustration of a gene with 10 SNPs constructed using alleles AA, Aa and aa.

		SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
(0)	AA Pr=maf^2	0	0	0	0	0	0	0	0	0	1
(1)	Aa Pr=$2 \cdot maf(1-mfa)$	0	0	0	0	0	1	0	0	0	0
(2)	aa Pr=$(1-maf)^2$	1	1	1	1	1	0	1	1	1	0
	x										

$x(2,2,2,2,2,1,2,2,2,0)$

Based on 1000 replications, we summarize the empirical bias, which is the mean difference of a true parameter and its estimates; the Mean Standard Error(MSE), which is the mean of asymptotic standard errors of parameter estimations; and the estimated standard deviation (ESD), which is the sample standard deviation of the estimates; and the confidence coverage probability (CP), which is the proportion of confidence intervals covering true parameter. These summary statistics of the simulation are represented in table 3.1.

Table 3.1: A summary of simulations based on Scenario I under the assumption that pairwise distance matrices of response and independent variables have a linear relationship–simple linear regression.

MAF	n	Empirical Bias		MSE		ESD		95% CI CP	
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
MAF=0.1	50	0.0002	0.000	0.0043	0.0006	0.0042	0.0006	1	1
	100	-0.0002	0.000	0.0031	0.0004	0.0031	0.0004	1	1
	200	-0.0000	0.000	0.0022	0.0003	0.0023	0.0003	1	1
MAF=0.3	50	0.0002	0.000	0.0048	0.0003	0.0047	0.0003	1	1
	100	0.0003	0.000	0.0033	0.0002	0.0034	0.0002	1	1
	200	0.0001	0.000	0.0023	0.0002	0.0023	0.0002	1	1
MAF=0.5	50	0.0002	0.000	0.0052	0.0003	0.0050	0.0003	1	1
	100	0.0001	0.000	0.0036	0.0002	0.0037	0.0002	1	1
	200	0.0001	0.000	0.0025	0.0002	0.0025	0.0001	1	1

In almost all cases the mean bias for each coefficient is zero with CI coverage proportion one. Mean of standard errors is almost equivalent to sample standard deviation of estimates. However, this scenario was trivial and in real world it's highly unlikely to observe accuracy to this degree. Next simulations are designed based on more complex relationships.

Scenario II: Similar to Scenario I, this simulation was set up imitate the genome structure a human gene, in this case with employing the frequency distribution of *MTR* gene [82] with $p = 20$ SNPs. A covariate z_i was introduced to model that is uniformly distributed on $(0, 1)$ and error term ε_i was drawn from normal distribution $N(0, .5^2)$. Use a true model: $y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_{ij} + \varepsilon_{ij}$, where, $\boldsymbol{\beta} = (0, 1, 1)'$ and $x_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|/(2p)$ (normalized Manhattan distance), $z_{ij} = (z_i^2 + z_j^2)^{1/2}$ and $\varepsilon_{ij} = \varepsilon_i - \varepsilon_j$, we simulated the response pairwise distances, y_{ij}. We also considered three different sample sizes, $n = 50, 100, 200$. Based on 1000 replications, we summarized the MSE, ESD, and CP of 95% confidence intervals in table Table 3.2.

We can see that with all the considered sample sizes, the empirical mean biases are small, and estimated standard errors are close to the mean standard errors. Besides, as sample size increases, mean bias decreases and mean of standard errors approaches sample standard deviation of estimates, and the coverage probability of 95% confidence intervals are about the nominal level. The results suggest that when the underlying model between the pairwise distance matrices is linear, our proposed estimators are consistent.

Table 3.2: A summary of simulations based on Scenario II under the assumption that pairwise distances are linearly related through a multiple linear regression model.

n	Empirical Bias		MSE		ESD		95% CI CP	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
50	0.0015	-0.001	0.365	0.095	0.363	0.088	0.95	0.95
200	-0.007	-0.001	0.176	0.041	0.174	0.041	0.95	0.95
400	-0.005	-0.0004	0.124	0.028	0.125	0.028	0.95	0.96

In our first two simulation scenarios, we make strong assumption on having a true linear relationship between the distance matrices. Considering the distance-based methods are also widely used in association studies about the relationship between

the original data, that is, between x and y. A considerable amount of literature has been published on the large sample theory for U-statistics, equivalency of distance covariances and distances between embedding of distributions to reproducing kernel Hilbert spaces (RKHS) and applications of kernel machine regression for test of association between the original x and y. [77, 79, 82–97] A question we asked to ourselves is "Can the pairwise distance based linear regression capture the underlying association of response and explanatory variables?" Another word to say it is if the true association is between the original variables, does the association retain in pairwise distance matrices?

Assume $(\mathbf{x}_i, \mathbf{y}_i), i = 1, 2, \cdots, n$, are n independent observations where $\mathbf{x}_i \in \mathcal{R}^p$ and $\mathbf{y}_i \in \mathcal{R}^q$. To test an association like $\mathbf{y}_i = \beta_0 + f(\mathbf{x}_i) + \varepsilon_i$ is proposed to set up the null hypothesis $H_0 : f(\cdot) = 0$. A useful type of test is a score test. Let

$$
\begin{aligned}
U &= (\mathbf{Y} - \bar{\mathbf{Y}})' K (\mathbf{Y} - \bar{\mathbf{Y}}) \\
&= tr[(\mathbf{Y} - \bar{\mathbf{Y}})' K (\mathbf{Y} - \bar{\mathbf{Y}})] \\
&= tr[K(\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}})] \\
&= tr\left[K\left(I - \frac{1_n 1_n'}{n}\right)\mathbf{Y}\mathbf{Y}'\left(I - \frac{1_n 1_n'}{n}\right)\right] \\
&= tr\left[KH\mathbf{Y}\mathbf{Y}'H\right]
\end{aligned}
$$

The outer product $\mathbf{Y}\mathbf{Y}'$ can be replaced with any distance matrix $D_{\mathbf{Y}}$ [88]. Let y_{ij} and x_{ij} be the ij^{th} elements of $D_{\mathbf{Y}}$ and $D_{\mathbf{X}}$ respectively. For if $y_{ij} = \alpha_0 + \alpha_1 \mathbf{x}_{ij} + \varepsilon_{ij}$, a choice of $D_{\mathbf{X}}$ (possibly any center-normalized distance matrix) and kernel $K = [k_{ij}]_{n \times n}$ satisfying (a) $0 \leq k_{ij} \leq 1$ and (b) $\mathbf{x}_{ij} = 1 - k_{ij}$ implies $y_{ij} = \alpha_0^* + \alpha_1^* k_{ij} + \varepsilon_{ij}$. Let

$U_\alpha = \sum_{1\le i<j\le n} k - ij(y_{ij} - \bar{d_Y})$, then

$$\begin{aligned} U_\alpha &= \frac{1}{2}\sum_{1\le i,j\le n} k_{ij}(y_{ij} - \bar{d_Y}) \\ &= \frac{1}{2}\left(\sum_{1\le i,j\le n} k_{ij}\cdot y_{ij} - \bar{d_Y}\sum_{1\le i,j\le n} k_{ij}\right) \\ &\equiv \frac{U}{2} - \frac{\bar{d_Y}}{2}\sum_{1\le i,j\le n} k_{ij} \end{aligned}$$

Hence, $U_\alpha \sim N(0, 4V_0)$ and consequently, the original hypothesis $H_0 : f(\cdot) = 0$ can be reduced to $H_0 : \alpha_1^* = 0$.

Simulations for Testing Hypotheses

Scenario III: This setup pertains to simulation study of kernel machine regression (KMR) by Hua and Ghosh [88]. Assuming $Y = c * h(\boldsymbol{X}) + \boldsymbol{Z}'\boldsymbol{\beta} + \varepsilon$ as true model when $h(.)$ is either $f_1(.) = g(.)$ or $f_2(.) = sign(g)|g(.)|^{1/2}$ and

$$g(X_i) = 1 + \sum_{k=1}^{p} X_{ik}\eta_k + \sum_{k=2}^{p} X_{i1}X_{ik}\gamma_k$$

with $\eta_1 = 0.4$, $\eta_2 = \cdots = \eta_p = 0.7$, $\gamma_2 = \cdots = \gamma_p = 0.2$. Quantifier c specified the departure from H_0 denoted as effect size. $X_{1\times p}$ was generated similar to Scenario II and covariate $\boldsymbol{Z}$ from standard bivariate normal distribution. The regression coefficients $\boldsymbol{\beta}$ =$(1, 1)$ and ε was drawn from standard normal, student's $t_{df=3}$ or $\chi^2_1 - 1$ distributions. For each sample, we first adjusted the effect of covariate using ordinary least squares regression and defined $\tilde{Y} = Y - \boldsymbol{Z}'\hat{\boldsymbol{\beta}}$. Distance based linear regression model in (Equation 3.1) in this case is $\tilde{y}_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}$, where x_{ij} is the pairwise Manhattan distance. Note that when $c = 0$, it corresponds to the null hypothesis of no effect of X. We set significance level $\alpha = 0.05$ to check the type-I error. To examine power, different c values were considered for f_1 (c = 0.1, 0.2, 0.3, 0.4, 0.5) and f_2 (c = 0.6, 1.2, 1.8, 2.4, 3) respectively. Quantifiers were assigned different range of values because the magnitudes of f_1 and f_2 are different. Results are shown in table Table 3.3. Empirical size is the proportion of null hypotheses rejected when null hypothesis was true and empirical power is the proportion of null hypotheses rejected when null hypothesis was in fact false. For both size and power of the test, the level of significance was set to $\alpha = 0.05$. Empirical Type-I error for all combinations of n, f, c lies within (-0.012,0.009) of 0.05 by which we learn false positive rate is not inflated. Empirical power is small for sample size $n = 50$ even with highest effect size and rapidly increases as sample size increases. Empirical power gradually increases

as effect size increases.

Table 3.3: $Y = c * f_j(\boldsymbol{X}) + \boldsymbol{Z}'\boldsymbol{\beta} + \varepsilon$; No sign of inflated false positive rate is observed. Power of test is sensitive to both sample and effect size.

				Empirical size	Empirical power				
h	ε	n	c	0	0.1	0.2	0.3	0.4	0.5
f_1	$N(0,1)$	50		0.055	0.06	0.104	0.19	0.297	0.409
		200		0.058	0.065	0.351	0.76	0.918	0.981
		400		0.054	0.121	0.667	0.972	0.998	1
f_1	$t_{df=3}$	50		0.038	0.055	0.061	0.103	0.183	0.225
		200		0.054	0.049	0.115	0.301	0.563	0.735
		400		0.048	0.056	0.192	0.525	0.78	0.95
f_1	$\chi^2_1 - 1$	50		0.05	0.054	0.081	0.144	0.188	0.287
		200		0.054	0.052	0.152	0.415	0.694	0.841
		400		0.059	0.085	0.275	0.709	0.988	0.988
			c	0	0.6	1.2	1.8	2.4	3
f_2	$N(0,1)$	50		0.047	0.066	0.1	0.134	0.206	0.304
		200		0.057	0.065	0.226	0.568	0.748	0.892
		400		0.057	0.095	0.437	0.887	0.983	0.998

Scenario IV: Following the line of Scenario III, for multivariate outcome, k = 1, 2 and 3, data were generated from the model $Y_k = c * h_k(\boldsymbol{X}) + \boldsymbol{Z}'\boldsymbol{\beta_k} + \varepsilon_k$ where $X_{1\times p}$ was generated using a frequency distribution of SNPs on the *SLC17A1* gene [88] with $p = 9$ SNPs. $\boldsymbol{Z}$ randomly sampled from bivariate normal distribution with means $\boldsymbol{\mu}$=$(0.2, 0.4)'$ and identity variance-covariance matrix. And, ε_k has a multivariate normal distribution MVNormal$(\mathbf{0}, \Sigma)$. Two choices for the effects of h_k were considered. First, the sparse effect, where

$$h_1 = c(x_1 + x_2 + x_3 + x_1x_4x_5 - x_6/3 - x_7x_8/2 + (1 - x_9)),$$

where $c = 0, 0.1, 0.2$ and $h_2 = h_3 = 0$.

Second, the common effect, where $h_1^* = h_1 + cx_3$, and $h_2 = h_3 = cx_3$ with $a = 0, 0.1, 0.2$. The variance–covariance matrix Σ was designed to have an independent structure ($\Sigma = \Sigma_1$) and a more dependent structure ($\Sigma = \Sigma_2$) as follows.

$$\Sigma_1 = \begin{bmatrix} 0.95 & 0 & 0 \\ 0 & 0.86 & 0 \\ 0 & 0 & 0.89 \end{bmatrix} \text{ and } \Sigma_2 = \begin{bmatrix} 0.95 & 0.57 & 0.43 \\ 0.57 & 0.86 & 0.24 \\ 0.43 & 0.24 & 0.89 \end{bmatrix}$$

Similar to scenario III, effect of covariate was adjusted using least squares regression, $\tilde{Y} = Y - \boldsymbol{Z}'\hat{\boldsymbol{\beta}}$. Then, the distance based linear regression model was carried out using theoretical model $\tilde{y}_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}$. Results are shown in table Table 3.4.

Table 3.4: Multivariate Outcome $Y_k = c * h_k(\boldsymbol{X}) + \boldsymbol{Z}'\boldsymbol{\beta_k} + \varepsilon_k$; Inflated rate of false negative is observed due to complex structure of true model.

				Empirical size	Empirical power	
h	Σ	n	c	0	0.1	0.2
$(h_1, 0, 0)$	Σ_1	50		0.041	0.044	0.088
		200		0.046	0.065	0.213
		400		0.052	0.087	0.452
$(h_1, 0, 0)$	Σ_2	50		0.044	0.038	0.067
		200		0.051	0.048	0.273
		400		0.048	0.077	0.476
			c	0	0.1	.2
(h_1^*, h_2, h_3)	Σ_1	50		0.046	0.05	0.098
		200		0.042	0.073	0.331
		400		0.054	0.119	0.633
(h_1^*, h_2, h_3)	Σ_2	50		0.05	0.043	0.08
		200		0.056	0.064	0.341
		400		0.044	0.081	0.572

Scenario V: The aim of this simulation study is to investigate the performance of DBLR in the world of compositional data. The structure of a data set X is referred as compositional when sample points comply with unit-sum constraint. Take for example ecological community data, e.g. relative abundance of microbial organisms per observation sums to one.

Aitchison [98] proposed to relax the unit-sum constraint by performing statistical analysis through log ratios. Among various forms of log-ratio transformations, the centered log-ratio transformation has attractive features and has been widely used. Based on [99], compositional data X was generated given $X_{ij}^{(k)} = W_{ij}^{(k)} / \sum_{l=1}^{p} W_{il}^{(k)}$, $i = 1, \ldots, n_k$, $j = 1, \ldots, p$, $k = 1, 2$ and $W_{n_k \times p}^{(k)} = (W_1^{(k)}, \ldots, W_{n_k}^{(k)})'$ denoting the matrices of unobserved bases. Let $Z_i^{(k)} = (Z_{i1}^{(k)}, \ldots, Z_{ip}^{(k)})$ be the log basis vectors, where $Z_{ij}^{(k)} = \log(W_{ij}^{(k)})$. For the observed compositional data $X^{(k)}$ $(k = 1, 2)$, the centered log-ratio matrices $Y^{(k)}$'s are defined by $Y_{ij}^{(k)} = \log\{X_{ij}^{(k)} / g(X_i^{(k)})\}$, $i = 1, \ldots, n_k$, $j = 1, \ldots, p$, $k = 1, 2$, where $g(x) = (\prod_{j=1}^{p} x_j)^{1/p}$ denotes the geometric mean of a vector x. Defining $G = I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p'$, this relation can be expressed as matrix form $Y_j^{(k)} = G \log(X_j^{(k)})$. According to scale-invariance property of the centered log-ratios, $X_j^{(k)}$ can be replaced by $W_j^{(k)}$ and obtain $Y_j^{(k)} = GZ_j^{(k)}$.

Simulation setup: Log basis vectors were generated from multivariate normal distribution, $Z_i^{(k)} \sim MVN_p(\mu_k, \Omega)$ $(k = 1, 2$ and $i = 1 \ldots n)$. The components of μ_1 were drawn from a uniform distribution $U(0, 10)$. Null and alternative hypotheses are considered as

$$\mathrm{H}_0 : \mu_2 = \mu_1 \text{ vs. } \mathrm{H}_a : \mu_{2j} = \mu_{1j} - \delta_j \omega_{jj}^{1/2} \left(\frac{\log p}{n} \right)^{1/2}$$

Then $W^{(k)}$ and $X^{(k)}$ were generated through the transformations $W_{ij}^{(k)} = \exp(Z_{ij}^{(k)})$ and $X_{ij}^{(k)} = W_{ij}^{(k)} / \sum_{l=1}^{p} W_{il}^{(k)}$. Signal vector δ has support of size $s = 0$, $\lfloor 0.05p \rfloor$, $\lfloor 0.1p \rfloor$ or $\lfloor 0.5p \rfloor$, p=50, 100, 200. Non-zero elements of δ were drawn from $Unif[-2\sqrt{2}, 2\sqrt{2}]$ with index chosen uniformly from $\{1, \ldots, p\}$. For covariance matrix, Ω was defined $\Omega = D^{1/2} A D^{1/2}$. D is a diagonal matrix with entries drawn from $Unif(1, 3)$ and A

has non-zero entries $a_{jj} = 1$ and $a_{j-1,j} = a_{j+1,j} = -0.5$.

Analysis: To assess the differential composition of two groups, first centered log-ratio transformation was applied to observations, $Y_j = clr(X_j)$ and then distance matrix D_Y was constructed for response variable. To adjust for the An indicator function was used to label elements of response as either within (w) or between groups (b). Using DBLR model

$$y_{ij} = \beta_0 I(\textit{samples i and j in group 1}) + \beta_1 I(\textit{at least one sample from group 2}) + \varepsilon_{ij}$$

and tested for $\mathrm{H}_0 : \beta_1 = 0$ vs $\mathrm{H}_1 : \beta_1 \neq 0$. Rejection of test concludes that there is shift in composition of group 2 regarding the composition of group 1; however it does not infer that dispersion of group 2 is different from group 1. See table Table 3.5 for empirical size and power of tests. Note two groups have same size $n_1 = n_2 = 100$ and significance level $\alpha = 0.05$.

Table 3.5: Empirical Size and Power of tests concerning the differential composition of groups. False positive rate if less than 1% and power increases as n or p increases.

s	$p = 50$	$p = 100$	$p = 200$
0	0.004	0.006	0.004
$\lfloor 0.05p \rfloor$	0.018	0.07	0.299
$\lfloor 0.1p \rfloor$	0.108	0.383	0.84
$\lfloor 0.5p \rfloor$	0.992	1	1

In continuation, two samples were generated in similar manner except $Z_i^{(1)}$ and $Z_i^{(2)}$ correlated and generated from a $2p$-dimensional joint distribution with mean $\mu^* = (\mu_1^{(k)T}, \mu_2^{(k)T})'$ and variance-covariance matrix

$$\Omega^* = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix} \otimes \Omega.$$

BBLR was fitted in similar to previous part and results are summarized in table Table 3.6. Empirical power follows the same trend as before. Moreover, power of the tests show less sensitivity to magnitude of p compared to uncorrelated design.

Table 3.6: Empirical Power for Samples with correlated mean structure.

s	$p = 50$	$p = 100$	$p = 200$
$\lfloor 0.05p \rfloor$	0.031	0.168	0.407
$\lfloor 0.1p \rfloor$	0.29	0.712	0.985
$\lfloor 0.5p \rfloor$	1	1	1

Scenario VI: In scope of analysis of differential composition, data was simulated to imitate community microbiome. Two group of samples were generated with identical or differential mean relative abundance, incorporating a negative binomial distribution with fixed parameters [100].

$$x_i \sim NB\left(x_i \middle| \mu_i, \theta_i\right) = \frac{\Gamma\left(x_i + \theta_i\right)}{\Gamma\left(\theta_i\right) x_i!} \cdot \left(\frac{\theta_i}{\mu_i + \theta}\right)^{\theta_i} \cdot \left(\frac{\mu_i}{\mu_i + \theta_i}\right)^{x_i} \tag{3.15}$$

where μ_i and $\phi_i = \frac{1}{\theta}$ are the mean and the dispersion parameter, respectively, and $\Gamma(.)$ is the gamma function. In the microbiome setting, μ_{ij} is considered as a product of the mean relative abundances $\rho_j = \frac{\sum_{i=1}^{n} x_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}}$ of taxon j and the library size $s_i = \sum_{j=1}^{p} x_{ij}$ of sample i, that is $\mu_{ij} = \rho_j s_i$. Library sizes were estimated using random subsets of taxonomic profiles of stool samples in HMP1 with replacement. Parameters ρ and θ were set to vectors of fixed values. Two groups of taxa ($p = 250$) tables with equal number of observations ($n_1 = n_2 = 50, 100, 200$) were simulated under the assumptions: (I) no differential abundance, i.e. fold changes were set to 1, (II) differential abundance by multiplying a fraction (true positive rate TPR = 0.25, 05, 0.75) of means and a fold change (FC = 1.5, 3, 5). Bray-Curtis distance [71], $BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$, is a commonly used distance measure in ecological data, where S_i is the total number of specimens counted on site i, S_j is the total number of

specimens counted on site j, and, C_{ij} is the sum of only the lesser counts for each specimen found on both sites. Bray-Curtis distances were computed and labeled as withing group 1, within group 2 and between groups. Considering the between-group distances as reference group, the DBLR model is

$$BC_{ij} = \beta_0 + \beta_1 I_{\text{within group 1}} + \beta_2 I_{\text{within group 2}} + \varepsilon_{ij} \tag{3.16}$$

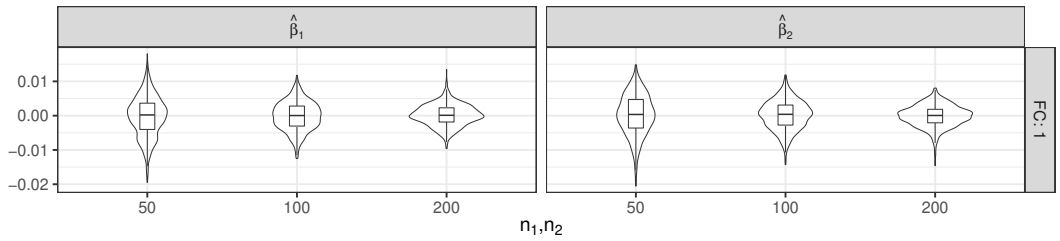
where $I(.)$ is the indicator function. Individual p-values were computed for each coefficient at significance level $\alpha = 0.05$. When $\beta_i > 0$ it can be interpreted as signal of additional variation of samples within that group to the mean distances of samples between two groups. When $\beta_i < 0$ it implies mean distances of samples withing that group have less dissimilarities compared to their dissimilarities to other sample. However, individual test cannot differentiate the cause whether it's due to a shift of centroid and/or scales of dispersion of groups. This might be investigated by multiple testing adjustment, testing the ratio of $\frac{\beta_1}{\beta_2}$, etc. Results are summarized in Table 3.7 and distribution of regression parameter estimates are visualized in Figure 3.2. There is no sign of inflated Type-I error. Empirical power seem to be more sensitive to sample size and less sensitive to TPR or FC. Power is low when $n = 50$ for any combination of TPR and FC. One might consider sample size estimation and power analysis.

Table 3.7: Summary of simulations for testing a difference of composition of microbiome of two groups. There is no sign of inflated Type-I error. Empirical power seem to be more sensitive to sample size and less sensitive to TPR or FC.

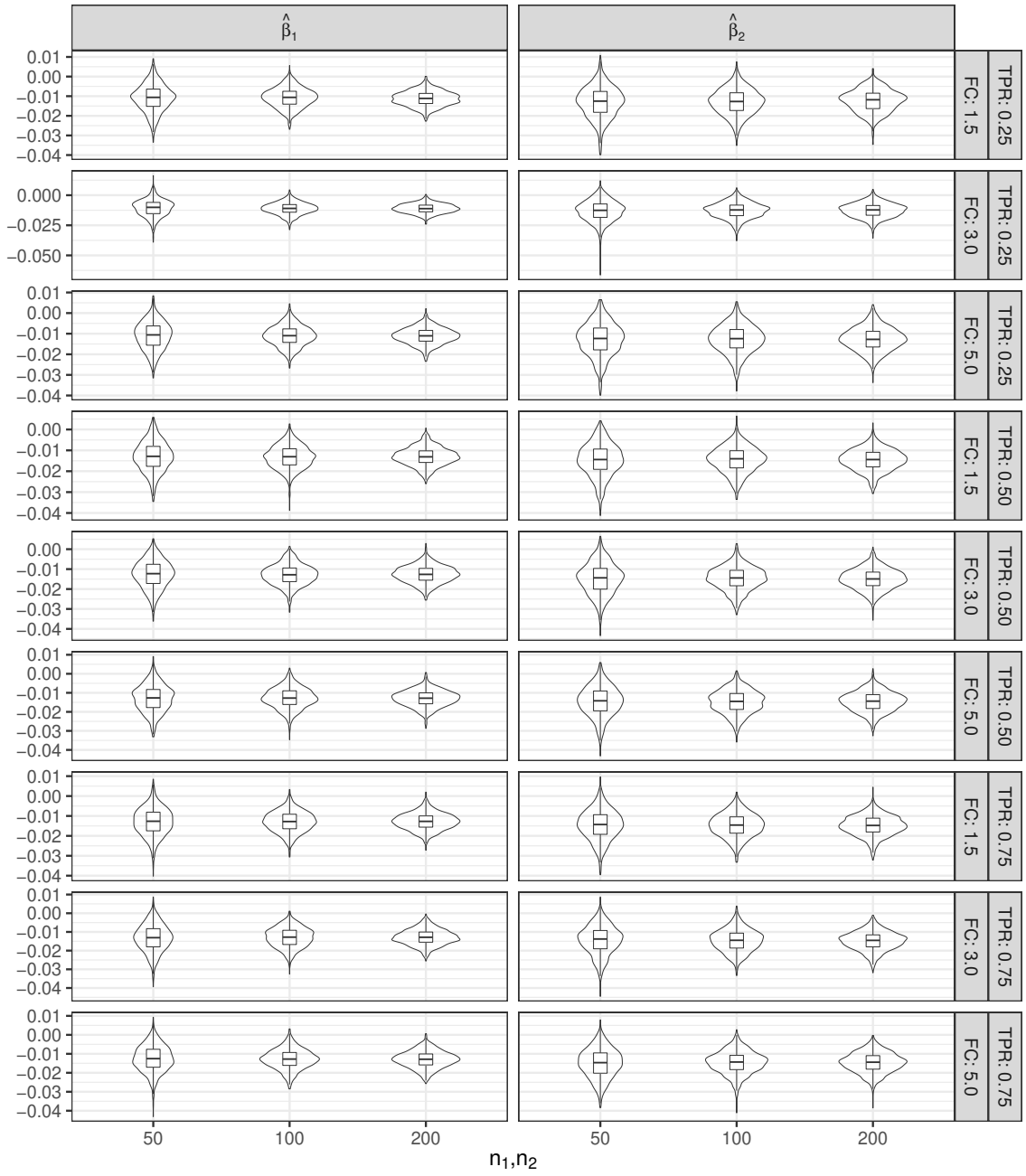
	n_1, n_2	Mean			MSE		ESD		P($p < 0.05$)	
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
FC=1	50	0.837	-0.000	0.000	0.006	0.006	0.006	0.006	0.047	0.049
	100	0.837	-0.000	0.000	0.004	0.004	0.004	0.004	0.048	0.046
	200	0.836	0.000	-0.000	0.003	0.003	0.003	0.003	0.053	0.053
TPR=0.25 FC=1.5	50	0.847	-0.011	-0.013	0.006	0.006	0.007	0.008	0.374	0.501
	100	0.848	-0.011	-0.013	0.004	0.005	0.005	0.007	0.670	0.742
	200	0.848	-0.011	-0.012	0.003	0.003	0.004	0.006	0.921	0.872
TPR=0.25 FC=3	50	0.847	-0.011	-0.013	0.006	0.007	0.007	0.009	0.370	0.502
	100	0.848	-0.011	-0.013	0.004	0.005	0.005	0.007	0.689	0.742
	200	0.848	-0.011	-0.013	0.003	0.003	0.004	0.006	0.916	0.879
TPR=0.25 FC=5	50	0.848	-0.011	-0.013	0.006	0.007	0.007	0.008	0.397	0.476
	100	0.847	-0.011	-0.013	0.004	0.005	0.005	0.007	0.689	0.728
	200	0.848	-0.011	-0.013	0.003	0.003	0.004	0.006	0.911	0.893
TPR=0.5 FC=1.5	50	0.850	-0.013	-0.015	0.006	0.007	0.007	0.008	0.529	0.596
	100	0.849	-0.013	-0.014	0.004	0.005	0.005	0.006	0.800	0.843
	200	0.849	-0.013	-0.015	0.003	0.003	0.004	0.005	0.956	0.971
TPR=0.5 FC=3	50	0.849	-0.013	-0.015	0.006	0.006	0.007	0.008	0.485	0.609
	100	0.849	-0.013	-0.015	0.004	0.005	0.005	0.006	0.819	0.871
	200	0.849	-0.013	-0.015	0.003	0.003	0.004	0.005	0.946	0.973
TPR=0.5 FC=5	50	0.850	-0.013	-0.014	0.006	0.006	0.007	0.008	0.517	0.590
	100	0.849	-0.013	-0.015	0.004	0.005	0.005	0.006	0.786	0.859
	200	0.849	-0.013	-0.015	0.003	0.003	0.004	0.005	0.964	0.966
TPR=0.75 FC=1.5	50	0.850	-0.013	-0.015	0.006	0.007	0.007	0.007	0.525	0.584
	100	0.849	-0.013	-0.015	0.004	0.005	0.005	0.006	0.800	0.859
	200	0.849	-0.013	-0.015	0.003	0.003	0.004	0.005	0.952	0.980
TPR=0.75 FC=3	50	0.849	-0.013	-0.014	0.006	0.006	0.007	0.008	0.531	0.578
	100	0.849	-0.013	-0.015	0.004	0.005	0.005	0.006	0.800	0.858
	200	0.849	-0.013	-0.015	0.003	0.003	0.004	0.005	0.948	0.974
TPR=0.75 FC=5	50	0.850	-0.013	-0.015	0.006	0.007	0.007	0.008	0.501	0.616
	100	0.849	-0.013	-0.015	0.004	0.005	0.005	0.006	0.801	0.866
	200	0.849	-0.013	-0.015	0.003	0.003	0.004	0.005	0.945	0.964

Figure 3.2: Box plots with added violin plots for visualizing the distribution of regression coefficients.

(a) Groups 1 and 2 independently simulated from identical distribution.



(b) Groups 1 and 2 independently simulated from when TPR·(100)% of relative mean abundance of group 2 is different from group 1 by multiplying a fold changes FC.



3.3 REAL DATA EXAMPLE

Analyses in this section pertains to HMP2 data introduced in Chapter 2. As previously studied using intensive simulations in Chapter 3, distance based linear regression can serve as a tool to generate hypotheses about the community data composition. Distances at species level were computed and labeled as withing and between UC, CD and nonIBD groups. Considering the within nonIBD distances as reference group, the DBLR model is

$$\begin{aligned} d_{ij} = \beta_0 &+ \beta_1 I_{\text{Within CD}} \\ &+ \beta_2 I_{\text{Between CD and UC}} \\ &+ \beta_3 I_{\text{Between CD and nonIBD}} \\ &+ \beta_4 I_{\text{ Within UC}} \\ &+ \beta_5 I_{\text{Between nonIBD and UC}} + \varepsilon_{ij} \end{aligned} \tag{3.17}$$

where $I(.)$ is the indicator function. Intercept parameter here is interpreted as mean distances of microbial composition of samples within the nonIBD group. Each other parameter is the difference of mean distances of other groups and nonIBD. Coefficient estimates, asymptotic standard errors and observed significance are summarized in Table 3.8 and Table 3.9. It's possible that DBLR has inflated Type-I error when B-C distance method is used.

Table 3.8: DBLR summary of estimations based on B-C distances of adult's gut microbiome species. Standard errors were computed via DBLR covariance estimation method.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.7061	0.0091	77.5403	0.0000
Within CD	0.0622	0.0135	4.6172	0.0000
Between CD and UC	0.0742	0.0116	6.4048	0.0000
Between CD and nonIBD	0.0552	0.0075	7.3336	0.0000
Within UC	0.0699	0.0133	5.2431	0.0000
Between nonIBD and UC	0.0632	0.0075	8.4236	0.0000

Table 3.9: DBLR summary of estimations based on WUnif distances of adult's gut microbiome species. Standard errors were computed via DBLR covariance estimation method.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.3174	0.0102	30.9843	0.0000
Within.CD	0.0525	0.0198	2.6473	0.0081
Between.CD.and.UC	0.0571	0.0160	3.5708	0.0004
Between.CD.and.nonIBD	0.0390	0.0097	4.0165	0.0001
Within.UC	0.0531	0.0198	2.6821	0.0073
Between.nonIBD.and.UC	0.0367	0.0099	3.7044	0.0002

CHAPTER 4: DISCUSSION

As mentioned in the literature review (Chapter 1), pairwise distance based statistical methods have been developed to study the functional relationship between these multivariate variables such as ecological community composition data. These methods fall into different categories such as distance based linear correlation association test e.g. Mantel test, multiple regression on distance matrices e.g. MRM, distance based multivariate analysis of variance e.g. PERMANOVA. However, a substantial amount of research argue that these methods suffer from inflated Type-I error and lack of power due to naive application of permutation testing.

In Chapter 2, after a brief review of "microbiome" is, we borrowed the metagenomic data from iHMP study and implemented some preliminary analysis and exemplified the applications of distance based multivariate techniques in microbiome data analysis.

In Chapter 3, we proposed a distance based linear regression model (DBLR) that is a linear regression on distance matrices. Estimates of parameters of DBLR are no different from ordinary least squares regression, but p-values are rather computed using our proposed method of estimation of variance-covariance matrix on a large sample theory basis. The performance of model was assessed through simulation studies and finished by an example of application of DBLR on HMP metagenomic data set.

Our method substantially reduces the needed computing power, since it does not rely on permutation. However, one of the challenges in our method is sensitivity to sample size considering asymptotic convergence of distribution. The connection between pairwise distance regression and regression model for raw data was explained by equivalency of kernel machine regression and kernel distance covariance. Research questions that could be asked include the the relationship between group centroids and dispersion, multiple testing adjustment, classification and model selection, sample size estimation and analysis of power.

REFERENCES

[1] H. Smith, R. Gnanadesikan, and J. B. Hughes, "Multivariate Analysis of Variance (MANOVA)," *International Biometric Society*, vol. 18, no. 1, pp. 22–41, 1962.

[2] R. F. Haase and M. V. Ellis, "Multivariate Analysis of Variance," *Journal of Counseling Psychology*, vol. 34, pp. 404–413, 10 1987.

[3] B. H. McArdle and M. J. Anderson, "Fitting multivariate models to community data: A comment on distance-based redundancy analysis," *Ecology*, vol. 82, no. 1, pp. 290–297, 2001.

[4] M. J. Anderson, "A new method for non-parametric multivariate analysis of variance," *Austral Ecology*, vol. 26, no. 1, pp. 32–46, 2001.

[5] N. MANTEL, "The Detection of Disease Clustering and a Generalized Regression Approach," *Cancer Research*, vol. 27, no. 2, pp. 209–220, 1967.

[6] E. J. Dietz, "Permutation Tests for Association between Two Distance Matrices," *Systematic Zoology*, vol. 32, no. 1, pp. 21–26, 1983.

[7] P. E. Smouse, J. C. Long, and R. R. Sokal, "Multiple Regression and Correlation Extensions of the Mantel Test of Matrix Correspondence," *Systematic Zoology*, vol. 35, no. 4, pp. 627–632, 1986.

[8] P. LEGENDRE and M.-J. FORTIN, "Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data," *Molecular Ecology Resources*, vol. 10, pp. 831–844, 9 2010.

[9] J. W. Lichstein, "Multiple regression on distance matrices: A multivariate spatial analysis tool," *Plant Ecology*, vol. 188, no. 2, pp. 117–131, 2007.

[10] H. Hotelling, "The Economics of Exhaustible Resources," *Journal of Political Economy*, vol. 39, 4 1931.

[11] S. S. Wilks, "Certain Generalizations in the Analysis of Variance," *Biometrika*, vol. 24, 11 1932.

[12] R. A. FISHER, "THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS," *Annals of Eugenics*, vol. 7, 9 1936.

[13] M. S. Bartlett, "A note on tests of significance in multivariate analysis," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 35, 4 1939.

[14] D. N. Lawley, "A Correction to A Generalization of Fisher's z Test"," *Biometrika*, vol. 30, 1 1939.

[15] K. C. S. Pillai, "Some New Test Criteria in Multivariate Analysis," Tech. Rep. 1, 1955.

[16] A. P. Dempster, "A High Dimensional Two Sample Significance Test," Tech. Rep. 4, 1958.

[17] A. P. Dempster, "A Significance Test for the Separation of Two Highly Multivariate Small Samples," Tech. Rep. 1, 1960.

[18] Y. Fujikoshi, T. Himeno, and H. Wakaki, "Asymptotic Results of a High Dimensional MANOVA Test and Power Comparison When the Dimension is Large Compared to the Sample Size," *JOURNAL OF THE JAPAN STATISTICAL SOCIETY*, vol. 34, no. 1, 2004.

[19] J. C. Gower, "A General Coefficient of Similarity and Some of Its Properties," *Biometrics*, vol. 27, p. 857, 12 1971.

[20] J. Gower, "Properties of Euclidean and non-Euclidean distance matrices," *Linear Algebra and its Applications*, vol. 67, 6 1985.

[21] J. C. Gower and P. Legendre, "Metric and Euclidean Properties of Dissimilarity Coefficients," tech. rep., 1986.

[22] J. C. Gower and W. J. Krzanowski, "Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance," *Journal of the Royal Statistical Society. Series C: Applied Statistics*, vol. 48, no. 4, pp. 505–519, 1999.

[23] M. J. Anderson and D. C. I. Walsh, "PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing?," Tech. Rep. 4, 2013.

[24] D. V. Lindley and M. R. Novick, "The Role of Exchangeability in Inference," *Source: The Annals of Statistics*, vol. 9, no. 1, pp. 45–58, 1981.

[25] M. J. Anderson, "Permutational Multivariate Analysis of Variance (PERMANOVA)," *Wiley StatsRef: Statistics Reference Online*, pp. 1–15, 2017.

[26] P. Legendre, M. Fortin, and D. Borcard, "Should the Mantel test be used in spatial analysis?," *Methods in Ecology and Evolution*, vol. 6, pp. 1239–1247, 11 2015.

[27] D. A. Griffith, "Spatial Autocorrelation," *Encyclopedia of Social Measurement*, pp. 581–590, 1 2004.

[28] F.-J. Lapointe and P. Legendre, "Comparison tests for dendrograms: A comparative evaluation," *Journal of Classification*, vol. 12, pp. 265–282, 9 1995.

[29] P. Legendre, "Comparison of permutation methods for the partial correlation and partial Mantel tests," *Journal of Statistical Computation and Simulation*, vol. 67, no. 1, pp. 37–73, 2000.

[30] L. J. Harmon and R. E. Glor, "Poor statistical performance of the mantel test in phylogenetic comparative analyses," *Evolution*, vol. 64, no. 7, pp. 2173–2178, 2010.

[31] D. I. Warton, S. T. Wright, and Y. Wang, "Distance-based multivariate analyses confound location and dispersion effects," *Methods in Ecology and Evolution*, vol. 3, pp. 89–101, 2 2012.

[32] S. M. Bloom, V. N. Bijanki, G. M. Nava, L. Sun, N. P. Malvin, D. L. Donermeyer, W. M. Dunne, P. M. Allen, and T. S. Stappenbeck, "Commensal Bacteroides Species Induce Colitis in Host-Genotype-Specific Fashion in a Mouse Model of Inflammatory Bowel Disease," *Cell Host & Microbe*, vol. 9, pp. 390–403, 5 2011.

[33] B. Hamidi, K. Wallace, C. Vasu, and A. V. Alekseyenko, "W_d* -test: Robust distance-based multivariate analysis of variance," *Microbiome*, vol. 7, no. 1, pp. 1–9, 2019.

[34] R. P. Franckowiak, M. Panasci, K. J. Jarvis, I. S. Acuña-Rodriguez, E. L. Landguth, M.-J. Fortin, and H. H. Wagner, "Model selection with multiple regression on distance matrices leads to incorrect inferences," *PLOS ONE*, vol. 12, p. e0175194, 4 2017.

[35] F. Zou, Z. Xu, and T. Vision, "Assessing the Significance of Quantitative Trait Loci in Replicable Mapping Populations," *Genetics*, vol. 174, pp. 1063–1068, 10 2006.

[36] G. A. Churchill and R. W. Doerge, "Naive Application of Permutation Testing Leads to Inflated Type I Error Rates," *Genetics*, vol. 178, pp. 609–610, 1 2008.

[37] Y. Huang, H. Xu, V. Calian, and J. C. Hsu, "To permute or not to permute," *Bioinformatics*, vol. 22, pp. 2244–2248, 9 2006.

[38] W. F. Christensen and B. N. Zabriskie, "When Your Permutation Test is Doomed to Fail," *The American Statistician*, pp. 1–11, 4 2021.

[39] L. K. Ursell, J. L. Metcalf, L. W. Parfrey, and R. Knight, "Defining the human microbiome," *Nutrition Reviews*, vol. 70, no. SUPPL. 1, 2012.

[40] J. Lederberg and A. McCray, "Ome Sweet Omics: a genealogical treasury of words," *The Scientist*, vol. 15, no. 7, pp. 8–8, 2001.

[41] J. Chen and D.-G. Chen, *Statistical Analysis of Microbiome Data with R - ICSA Book Series in Statistics*. Springer US, 2018.

[42] G. D. Wu, F. D. Bushmanc, and J. D. Lewis, "Diet, the human gut microbiota, and IBD," *Anaerobe*, vol. 24, pp. 117–120, 2013.

[43] E. Bianconi, A. Piovesan, F. Facchin, A. Beraudi, R. Casadei, F. Frabetti, L. Vitale, M. C. Pelleri, S. Tassani, F. Piva, S. Perez-Amodio, P. Strippoli, and S. Canaider, "An estimation of the number of cells in the human body," *Annals of Human Biology*, vol. 40, pp. 463–471, 11 2013.

[44] R. D. Berg, "The indigenous gastrointestinal microflora," 11 1996.

[45] D. C. Savage, "Microbial Ecology of the Gastrointestinal Tract," *Annual Review of Microbiology*, vol. 31, pp. 107–133, 10 1977.

[46] R. Sender, S. Fuchs, and R. Milo, "Revised Estimates for the Number of Human and Bacteria Cells in the Body," *PLOS Biology*, vol. 14, p. e1002533, 8 2016.

[47] G. G. Kaplan, "The global burden of IBD: from 2015 to 2025," *Nature Reviews Gastroenterology & Hepatology*, vol. 12, pp. 720–727, 12 2015.

[48] J. R. Kelsen and G. D. Wu, "The gut microbiota, environment and diseases of modern society," *Gut Microbes*, vol. 3, pp. 374–382, 7 2012.

[49] R. B. Sartor, "Therapeutic correction of bacterial dysbiosis discovered by molecular techniques," *Proceedings of the National Academy of Sciences*, vol. 105, pp. 16413–16414, 10 2008.

[50] T. Integrative, "The integrative human microbiome project: Dynamic analysis of microbiome-host omics profiles during periods of human health and disease corresponding author," *Cell Host and Microbe*, vol. 16, no. 3, pp. 276–289, 2014.

[51] L. M. Proctor, H. H. Creasy, J. M. Fettweis, J. Lloyd-Price, A. Mahurkar, W. Zhou, G. A. Buck, M. P. Snyder, J. F. Strauss, G. M. Weinstock, O. White, and C. Huttenhower, "The Integrative Human Microbiome Project," *Nature*, vol. 569, pp. 641–648, 5 2019.

[52] J. M. Fettweis, M. G. Serrano, J. L. P. L. Brooks, D. J. Edwards, P. H. Girerd, H. I. Parikh, B. Huang, T. J. Arodz, L. Edupuganti, A. L. Glascock, J. Xu, N. R. Jimenez, S. C. Vivadelli, S. S. Fong, N. U. Sheth, S. Jean, V. Lee, Y. A. Bokhari, A. M. Lara, S. D. Mistry, R. A. Duckworth, S. P. Bradley, V. N. Koparde, X. V. Orenda, S. H. Milton, S. K. Rozycki, A. V. Matveyev, M. L. Wright, S. V. Huzurbazar, E. M. Jackson, E. Smirnova, J. Korlach, Y.-C. Tsai, M. R. Dickinson, J. L. P. L. Brooks, J. I. Drake, D. O. Chaffin, A. L. Sexton, M. G. Gravett, C. E. Rubens, N. R. Wijesooriya, K. D. Hendricks-Muñoz, K. K. Jefferson, J. F. Strauss, and G. A. Buck, "The vaginal microbiome and preterm birth," *Nature Medicine*, vol. 25, pp. 1012–1021, 6 2019.

[53] J. Lloyd-Price, C. Arze, A. N. Ananthakrishnan, M. Schirmer, J. Avila-Pacheco, T. W. Poon, E. Andrews, N. J. Ajami, K. S. Bonham, C. J. Brislawn, D. Casero,

H. Courtney, A. Gonzalez, T. G. Graeber, A. B. Hall, K. Lake, C. J. Landers, H. Mallick, D. R. Plichta, M. Prasad, G. Rahnavard, J. Sauk, D. Shungin, Y. Vázquez-Baeza, R. A. White, J. Braun, L. A. Denson, J. K. Jansson, R. Knight, S. Kugathasan, D. P. B. McGovern, J. F. Petrosino, T. S. Stappenbeck, H. S. Winter, C. B. Clish, E. A. Franzosa, H. Vlamakis, R. J. Xavier, C. Huttenhower, J. Bishai, K. Bullock, A. Deik, C. Dennis, J. L. Kaplan, H. Khalili, L. J. McIver, C. J. Moran, L. Nguyen, K. A. Pierce, R. Schwager, A. Sirota-Madi, B. W. Stevens, W. Tan, J. J. ten Hoeve, G. Weingart, R. G. Wilson, V. Yajnik, J. Braun, L. A. Denson, J. K. Jansson, R. Knight, S. Kugathasan, D. P. B. McGovern, J. F. Petrosino, T. S. Stappenbeck, H. S. Winter, C. B. Clish, E. A. Franzosa, H. Vlamakis, R. J. Xavier, and C. Huttenhower, "Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases," *Nature*, vol. 569, pp. 655–662, 5 2019.

[54] W. Zhou, M. R. Sailani, K. Contrepois, Y. Zhou, S. Ahadi, S. R. Leopold, M. J. Zhang, V. Rao, M. Avina, T. Mishra, J. Johnson, B. Lee-McMullen, S. Chen, A. A. Metwally, T. D. B. Tran, H. Nguyen, X. Zhou, B. Albright, B.-Y. Hong, L. Petersen, E. Bautista, B. Hanson, L. Chen, D. Spakowicz, A. Bahmani, D. Salins, B. Leopold, M. Ashland, O. Dagan-Rosenfeld, S. Rego, P. Limcaoco, E. Colbert, C. Allister, D. Perelman, C. Craig, E. Wei, H. Chaib, D. Hornburg, J. Dunn, L. Liang, S. M. S.-F. Rose, K. Kukurba, B. Piening, H. Rost, D. Tse, T. McLaughlin, E. Sodergren, G. M. Weinstock, and M. Snyder, "Longitudinal multi-omics of host–microbe dynamics in prediabetes," *Nature*, vol. 569, pp. 663–671, 5 2019.

[55] H. Mallick, E. A. Franzosa, L. J. Mclver, S. Banerjee, A. Sirota-Madi, A. D. Kostic, C. B. Clish, H. Vlamakis, R. J. Xavier, and C. Huttenhower, "Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences," *Nature Communications*, vol. 10, p. 3136, 12 2019.

[56] E. A. Franzosa, L. J. McIver, G. Rahnavard, L. R. Thompson, M. Schirmer, G. Weingart, K. S. Lipson, R. Knight, J. G. Caporaso, N. Segata, and C. Huttenhower, "Species-level functional profiling of metagenomes and metatranscriptomes," *Nature Methods*, vol. 15, pp. 962–968, 11 2018.

[57] L. J. McIver, G. Abu-Ali, E. A. Franzosa, R. Schwager, X. C. Morgan, L. Waldron, N. Segata, and C. Huttenhower, "bioBakery: a meta'omic analysis environment," *Bioinformatics*, vol. 34, pp. 1235–1237, 4 2018.

[58] M. Schirmer, E. A. Franzosa, J. Lloyd-Price, L. J. McIver, R. Schwager, T. W. Poon, A. N. Ananthakrishnan, E. Andrews, G. Barron, K. Lake, M. Prasad, J. Sauk, B. Stevens, R. G. Wilson, J. Braun, L. A. Denson, S. Kugathasan, D. P. B. McGovern, H. Vlamakis, R. J. Xavier, and C. Huttenhower, "Dynamics of metatranscription in the inflammatory bowel disease gut microbiome," *Nature Microbiology*, vol. 3, pp. 337–346, 3 2018.

[59] E. A. Franzosa, A. Sirota-Madi, J. Avila-Pacheco, N. Fornelos, H. J. Haiser, S. Reinker, T. Vatanen, A. B. Hall, H. Mallick, L. J. McIver, J. S. Sauk, R. G. Wilson, B. W. Stevens, J. M. Scott, K. Pierce, A. A. Deik, K. Bullock, F. Imhann, J. A. Porter, A. Zhernakova, J. Fu, R. K. Weersma, C. Wijmenga, C. B. Clish, H. Vlamakis, C. Huttenhower, and R. J. Xavier, "Gut microbiome structure and metabolic activity in inflammatory bowel disease," *Nature Microbiology*, vol. 4, pp. 293–305, 2 2019.

[60] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower, "Metagenomic microbial community profiling using unique clade-specific marker genes," *Nature Methods*, vol. 9, pp. 811–814, 8 2012.

[61] R. Knight, A. Vrbanac, B. C. Taylor, A. Aksenov, C. Callewaert, J. Debelius, A. Gonzalez, T. Kosciolek, L.-I. McCall, D. McDonald, A. V. Melnik, J. T. Morton, J. Navas, R. A. Quinn, J. G. Sanders, A. D. Swafford, L. R. Thompson, A. Tripathi, Z. Z. Xu, J. R. Zaneveld, Q. Zhu, J. G. Caporaso, and P. C. Dorrestein, "Best practices for analysing microbiomes," *Nature Reviews Microbiology*, vol. 16, 7 2018.

[62] R. H. Whittaker, "EVOLUTION AND MEASUREMENT OF SPECIES DIVERSITY," *TAXON*, vol. 21, pp. 213–251, 5 1972.

[63] T. M. DeJong, "A Comparison of Three Diversity Indices Based on Their Components of Richness and Evenness," *Oikos*, vol. 26, no. 2, p. 222, 1975.

[64] C. H. Heip, P. M. J. Herman, and K. Soetaert, "Indices of diversity and evenness Project DISCLOSE View project Respiration in ocean margin sediments View project," *Oceanis*, vol. 24, no. 4, pp. 61–87, 1998.

[65] J. J. Sepkoski, "Alpha, beta, or gamma: where does all the diversity go?," *Paleobiology*, vol. 14, pp. 221–234, 2 1988.

[66] G. Stirling and B. Wilsey, "Empirical Relationships between Species Richness, Evenness, and Proportional Diversity," *the american naturalist*, vol. 158, no. 3, pp. 286–299, 2001.

[67] N. J. Gotelli and R. K. Colwell, "Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness," *Ecology Letters*, vol. 4, 7 2001.

[68] L. Jost, "The Relation between Evenness and Diversity," *Diversity*, vol. 2, pp. 207–232, 2 2010.

[69] K. Li, M. Bihan, S. Yooseph, and B. A. Methé, "Analyses of the Microbial Diversity across the Human Microbiome," *PLoS ONE*, vol. 7, p. e32118, 6 2012.

[70] W. H. Kruskal and W. A. Wallis, "Use of Ranks in One-Criterion Variance Analysis," *Journal of the American Statistical Association*, vol. 47, 12 1952.

[71] J. R. Bray and J. T. Curtis, "An Ordination of the Upland Forest Communities of Southern Wisconsin," *Ecological Monographs*, vol. 27, 10 1957.

[72] C. Lozupone and R. Knight, "UniFrac: a New Phylogenetic Method for Comparing Microbial Communities," *APPLIED AND ENVIRONMENTAL MICROBIOLOGY*, vol. 71, no. 12, pp. 8228–8235, 2005.

[73] C. A. Lozupone, M. Hamady, S. T. Kelley, and R. Knight, "Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities," *Applied and Environmental Microbiology*, vol. 73, pp. 1576–1585, 3 2007.

[74] M. Hamady, C. Lozupone, and R. Knight, "Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data," *The ISME Journal*, vol. 4, pp. 17–27, 1 2010.

[75] C. Lozupone, M. E. Lladser, D. Knights, J. Stombaugh, and R. Knight, "UniFrac: an effective distance metric for microbial community comparison," *The ISME Journal*, vol. 5, pp. 169–172, 2 2011.

[76] J. C. Gower, "Principal Coordinates Analysis," in *Wiley StatsRef: Statistics Reference Online*, Wiley, 3 2015.

[77] W. Hoeffding, "A Class of Statistics with Asymptotically Normal Distribution," *The Annals of Mathematical Statistics*, vol. 19, pp. 293–325, 9 1948.

[78] J. N. Arvesen, "Jackknifing U-Statistics," *The Annals of Mathematical Statistics*, vol. 40, no. 6, pp. 2076–2100, 1969.

[79] G. G. Gregory, "Large Sample Theory for U-Statistics and Tests of Fit," *The Annals of Statistics*, vol. 5, no. 1, pp. 110–123, 1977.

[80] S. Clémençon, G. Lugosi, and N. Vayatis, "Ranking and empirical minimization of U-statistics," *Annals of Statistics*, vol. 36, no. 2, pp. 844–874, 2008.

[81] B. E. Honoré and J. L. Powell, "Pairwise difference estimators of censored and truncated regression models," *Journal of Econometrics*, vol. 64, no. 1-2, pp. 241–278, 1994.

[82] D. Kong, A. Maity, F. C. Hsu, and J. Y. Tzeng, "Testing and estimation in marker-set association study using semiparametric quantile regression kernel machine," *Biometrics*, vol. 72, no. 2, pp. 364–371, 2016.

[83] H. Callaert and P. Janssen, "The Berry-Esseen Theorem for U-Statistics," *The Annals of Statistics*, vol. 6, no. 2, pp. 417–421, 1978.

[84] B. E. Honoré and J. L. Powell, "Pairwise Difference Estimators for Nonlinear Models," in *Identification and Inference for Econometric Models*, pp. 520–553, Cambridge University Press, 6 2005.

[85] P. Kumar Sen, "Robust Statistical Inference for High-Dimensional Data Models with Application to Genomics," Tech. Rep. 2&3, 2006.

[86] P. S. Zhong and S. X. Chen, "Tests for high-dimensional regression coefficients with factorial designs," *Journal of the American Statistical Association*, vol. 106, no. 493, pp. 260–274, 2011.

[87] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu, "EQUIVALENCE OF DISTANCE-BASED AND RKHS-BASED STATISTICS IN HYPOTHESIS TESTING," *The Annals of Statistics*, vol. 41, no. 5, pp. 2263–2291, 2013.

[88] W. Y. Hua and D. Ghosh, "Equivalence of kernel machine regression and kernel distance covariance for multidimensional phenotype association studies," *Biometrics*, vol. 71, no. 3, pp. 812–820, 2015.

[89] Q. He, T. Cai, Y. Liu, N. Zhao, Q. E. Harmon, L. M. Almli, E. B. Binder, S. M. Engel, K. J. Ressler, K. N. Conneely, X. Lin, and M. C. Wu, "Prioritizing individual genetic variants after kernel machine testing using variable selection," *Genetic Epidemiology*, vol. 40, no. 8, pp. 722–731, 2016.

[90] Z. Z. Tang, G. Chen, and A. V. Alekseyenko, "PERMANOVA-S: Association test for microbial community composition that accommodates confounders and multiple distances," *Bioinformatics*, vol. 32, no. 17, pp. 2618–2625, 2016.

[91] O. Paliy and V. Shankar, "Application of multivariate statistical techniques in microbial ecology," *Molecular Ecology*, vol. 25, no. 5, pp. 1032–1057, 2016.

[92] X. Zhan, X. Tong, N. Zhao, A. Maity, M. C. Wu, and J. Chen, "A small-sample multivariate kernel machine test for microbiome association studies," *Genetic Epidemiology*, vol. 41, no. 3, pp. 210–220, 2017.

[93] H. Koh, M. J. Blaser, and H. Li, "A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping," *Microbiome*, vol. 5, no. 1, pp. 1–15, 2017.

[94] M. Luz Calle, "Statistical analysis of metagenomics data," *Genomics and Informatics*, vol. 17, no. 1, 2019.

[95] T. He, S. Li, P. S. Zhong, and Y. Cui, "An optimal kernel-based U-statistic method for quantitative gene-set association analysis," *Genetic Epidemiology*, vol. 43, no. 2, pp. 137–149, 2019.

[96] N. B. Larson, J. Chen, and D. J. Schaid, "A review of kernel methods for genetic association studies," *Genetic Epidemiology*, vol. 43, pp. 122–136, 3 2019.

[97] M. Fuchs, R. Hornung, A. L. Boulesteix, and R. De Bin, "On the asymptotic behaviour of the variance estimator of a U-statistic," *Journal of Statistical Planning and Inference*, vol. 209, pp. 101–111, 2020.

[98] J. Aitchison, "The Statistical Analysis of Compositional Data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 44, 1 1982.

[99] Y. Cao, W. Lin, and H. Li, "Two-sample tests of high-dimensional means for compositional data," *Biometrika*, vol. 105, no. 1, pp. 115–132, 2018.

[100] S. Hawinkel, F. Mattiello, L. Bijnens, and O. Thas, "A broken promise: microbiome differential abundance methods do not control the false discovery rate," *Briefings in Bioinformatics*, vol. 20, pp. 210–221, 1 2019.